AD_____

Award Number: W81XWH-05-1-0292

TITLE: A Computer-Aided Diagnosis System for Breast Cancer Combining Digital Mammography and Genomics

PRINCIPAL INVESTIGATOR: Jonathan Jesneck
Joseph Lo, Ph.D.

CONTRACTING ORGANIZATION: Duke University
Durham, NC 27710

REPORT DATE: May 2006

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-05-2006	Annual Summary	1 MAY 2005 - 30 APR 2006

4. TITLE AND SUBTITLE
A Computer-Aided Diagnosis System for Breast Cancer Combining Digital Mammography and Genomics

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0292

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jonathan Jesneck
Joseph Lo, Ph.D.

E-mail: jonathan.jesneck@duke.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University
Durham, NC 27710

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains color plates: All DTIC reproductions will be in black and white.

14. ABSTRACT

This study investigated a computer-aided diagnosis system for breast cancer by combining the following three data sources: mammogram films, radiologist-interpreted BI-RADS descriptors, and proteomic profiles of blood sera. In this first year of the fellowship, we have collected calcification and mass data sets. To these data sets we have applied the following classification algorithms: Bayesian probit regression, linear discriminant analysis, artificial neural networks, as well as a novel method of decision fusion. For the calcification data set, the classifiers' performances under 100-fold cross validation were AUC = 0.73 for Bayesian probit regression, 0.68 ± 0.01 for LDA, 0.76 ± 0.01 for ANN, 0.85 ± 0.01 for decision fusion . For the mass data set, the classifiers' performances under 100-fold cross validation were AUC = 0.94 for Bayesian probit regression, 0.93 ± 0.01 for LDA, 0.93 ± 0.01for ANN, 0.94 ± 0.01 for decision fusion. Decision fusion had a slight performance gain over the ANN and LDA ($p = 0.02$), but was comparable to Bayesian probit regression. Decision fusion significantly outperformed the other classifiers ($p < 0.001$).

15. SUBJECT TERMS
computer-aided diagnosis, digital mammography, clinical proteomics, biopsy, receiver operating characteristic, Bayesian regression, ensemble classifier

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					USAMRMC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	UU	88	

Table of Contents

Cover…….. 1

Table of Contents…….. 2

SF 298…….. 3

Introduction…….. 4

Body…….. 4

Key Research Accomplishments…….. 6

Reportable Outcomes…….. 7

Conclusions…….. 7

References…….. 7

Appendices…….. 8

INTRODUCTION

This study investigated a computer-aided diagnosis (CADx) system for breast cancer by combining the following three data sources: mammogram films, radiologist-interpreted BI-RADS descriptors, and proteomic profiles of blood sera.

Although mammography is the modality of choice for early detection of breast cancer[1,2], it has a low positive predictive value (PPV). As a result, only 15 to 34% of women with radiographically-suspicious, nonpalpable lesions are actually found to have a malignancy by histologic diagnosis after biopsy.[3,4] The excessive biopsy of benign lesions raises the cost of mammographic screening[5] and results in emotional and physical burden to the patients, as well as financial burden to society.

In addition to mammography, both BI-RADS descriptors[5] and clinical proteomics[6] have been useful in differentiating benign from malignant breast masses. The combination of mammographic and proteomic information can lead to a more specific classifier for difficult cases. Ensemble classifiers for breast cancer combining multiple sources of information have been shown to outperform classifiers using only one of the information sources.[7]

This research has two purposes. The first is to create three separate classifiers for breast cancer based on proteomic information, mammogram information, and radiologist-interpreted. The second is to combine the outputs of these three first-stage classifiers into one ensemble classifier for breast cancer, which will outperform any of the component classifiers.

BODY

Task 1. Build a Bayesian regression model classifier for breast cancer based on image features of digitized mammograms. Evaluate the model performance using honest leave-one-out cross-validation (LOOCV) with the ROC area as the performance metric. Calculate the Bayesian posterior classification probability intervals to provide an honest assessment of the uncertainties of the predictive classifications. (Months 1-12)

This task has already been completed during the current, first year and has resulted in one accepted and one submitted peer-reviewed publication as well as one full-length conference proceedings paper (see #1, #2, and #3 in Reportable Outcomes). On each digitized mammogram, a 512x512 region of interest (ROI) centered on the centroid of each calcification cluster was extracted. The automated image-processing scheme consisted of the following steps: (1) pre-processing using unsharp masking, (2) segmentation of individual calcifications using a back-propagation artificial neural network (BP-ANN) classifier, and (3) cluster classification using another BP-ANN classifier to reduce the number of false positive clusters. For each cluster, the algorithm calculated 22 image-processing features, consisting mostly of shape features for the calcifications and calcification clusters and of texture features for ROIs centered on the clusters.

Once the features had been extracted from the mammogram, they were used to distinguish benign from malignant calcification lesions by classification models. In addition to Bayesian probit regression models, for comparison we also applied two well-established CADx classifiers, linear discriminant analysis (LDA), artificial neural network (ANN). We also applied two variants of a novel classifier, decision fusion: decision fusion to maximize the area under the ROC curve (DF-A), and to maximize the high-sensitivity region (TPF $\geq$ 0.90) partial area (DF-P). Decision fusion was a novel classification method (See #1 in Reportable Outcomes). Figure 1a shows the ROC curve for the Bayesian probit regression, and Figure 1b shows the set of ROC curves for the classifiers' performances under 100-fold cross validation were AUC = 0.73 for

Bayesian probit regression, 0.68 ± 0.01 for LDA, 0.76 ± 0.01 for ANN, 0.85 ± 0.01 for DF-A, and 0.82 ± 0.01 for DF-P. Decision fusion significantly outperformed the other classifiers ($p < 0.001$).

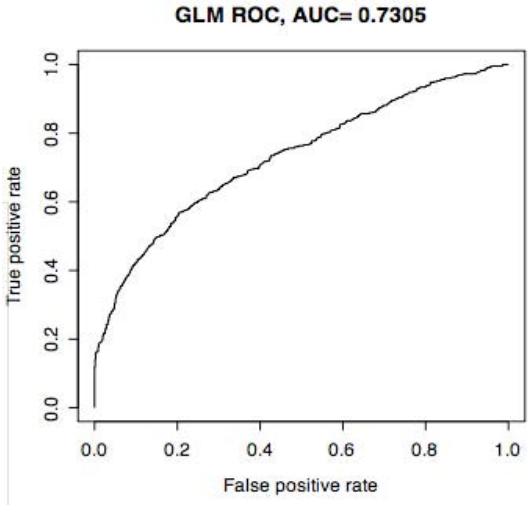


Figure 1a: Bayesian probit regression

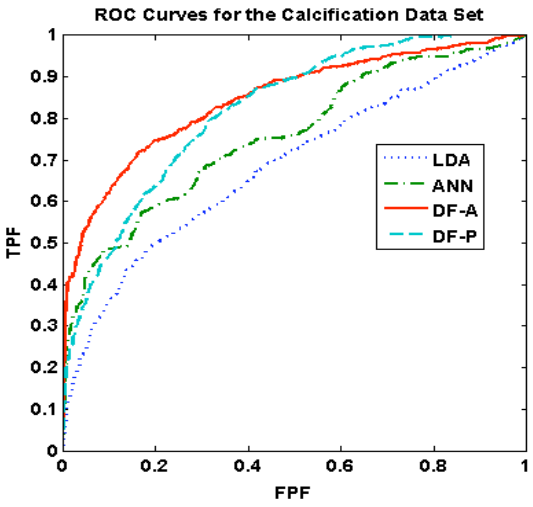


Figure 1b: LDA, ANN, and decision fusion

Task 2. Build a Bayesian regression model classifier for breast cancer based patient age and BI-RADS features from radiologists. Evaluate the model performance and classification uncertainties as in Aim 1. (Months 13-16)

This task has already been completed and has resulted in publications (see #1 and #2 in Reportable Outcomes). The mammographic findings for each case in our database have been interpreted by dedicated breast imaging radiologists using the Breast Imaging Reporting and Data System (BI-RADS) lexicon from the American College of Radiology.[8] The BI-RADS lexicon provides categorical descriptions (findings) for each mammographic feature.

While the original research proposal focused only on microcalcification lesions, we have responded to one of the proposal reviewers and have extended the research project to include masses as well. Including masses will lend additional clinical relevance to this project. Currently, the radiologist-interpreted BI-RADS features are available only for mass cases.

All of the classifiers were able to distinguish benign from malignant lesions well. The classifiers' performances under 100-fold cross validation were AUC = 0.94 for Bayesian probit regression, 0.93 ± 0.01 for LDA, 0.93 ± 0.01for ANN, 0.94 ± 0.01for DF-A, and 0.93 ± 0.01 for DF-P. Decision fusion had a slight performance gain over the ANN and LDA ($p = 0.02$), but was comparable to Bayesian probit regression. The ROC curves of these classifiers are shown in Figures 2a and 2b.

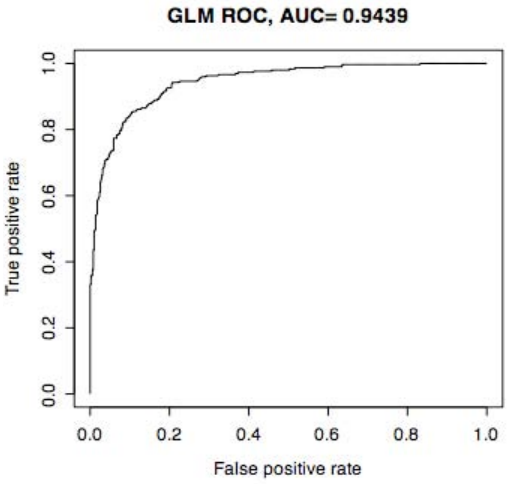


Figure 2a: Bayesian probit regression

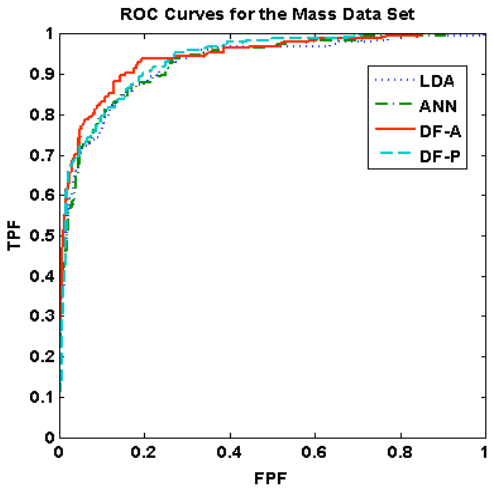


Figure 2b: LDA, ANN, and decision fusion

Task 3. Build a Bayesian regression model classifier for breast cancer based on proteomic profiles of blood serum samples. Evaluate the model performance and classification uncertainties as in Aim 1. (Months 16-28)

We have done some preliminary work on the proteomics data, and our research here is still a work in progress.

Women undergoing diagnostic biopsy at Duke University Medical Center for breast cancer between 2000-2004 were enrolled in this study. Before cytoreductive surgery, women were consented for the study and blood was obtained. Serum, plasma, and white blood cells were aliquoted and cryogenically stored. Two sets were constructed from these samples: 1) Forty-two women over the age of 55 with benign breast findings and 2) Forty-six women over the age of 55 with invasive breast cancers greater than 1.5 cm. In addition, sera from 120 healthy women were used for controls.

While the original research proposal included proteomic data from mass spectrometry spectra, these spectra were found to be too noisy for the purposes of classifying malignant from benign lesions. We are now using the much more specific Enzyme-Linked ImmunoSorbent Assay (ELISA) protocol to extract information about blood serum proteins. Sera were assayed for 52 different biomarkers using the Luminex platform and reagents. Because these biomarkers are expensive to collect, we are currently trying to identify a subset of important proteins by exploring feature-selection techniques on our proteomics pilot data set. Once the important proteins have been identified, more cases will be collected, allowing for further modeling and classifying.

Task 4. Combine the outputs of the three Bayesian regression models into one ensemble classifier for breast cancer diagnosis prediction. Evaluate the model performance using the ROC area as the performance metric. (Months 28-36)

Once we have finalized all three of the separate models described above, we will combine them into one ensemble classifier.

KEY RESEARCH ACCOMPLISHMENTS

- Developed a decision fusion model to combine various information sources

- Classified the mammogram and BI-RADS data sets using the following classification models: Bayesian probit regression, linear discriminant analysis, artificial neural network, and decision fusion
- Established an internal collaboration as a data source for the proteomics data set, and initiated preliminary analysis of that data set.

CONCLUSIONS

The current work focuses on combining breast imaging and proteomics information for breast cancer diagnosis. This study is structured in two stages: (1) build classification models on each of the individual data sources, and (2) combine the models into one ensemble classifier.

One significant research outcome was the development of a decision fusion classification algorithm. Decision fusion has the benefit of being robust in very noisy data sets, such as the calcification and proteomics data sets. On the more challenging calcification data set, decision fusion outperformed the other classifiers by achieving AUC = 0.85 ± 0.01. On the BI-RADS data set, all classifiers performed well, with decision fusion still performing the best with AUC = 0.94 ± 0.01.

The proteomics work is still a work in progress, due to the relatively small number of cases that are currently available as well as the large number of noisy features in the data set. In future work, we will identify a subset of blood serum proteins that are useful for breast cancer classification. Once these proteins have been identified, we will collect more cases to increase the size of the proteomics data set. With a larger data set, we can construct predictive models. Finally, once these models for all three data sets have been finalized, we will combine them into one ensemble classifier.

REPORTABLE OUTCOMES

The following publications are attached as appendices 1-4 with the same numbers. The names of the fellow (Jesneck) and mentor (Lo) are boldfaced for emphasis.

1 **Jesneck JL**, Nolte LW, Baker JA, Floyd CE, **Lo JY**, "An optimized approach to decision fusion of heterogeneous data for Breast Cancer Diagnosis," *Medical Physics*, (in press)

2 **Jesneck JL**, **Lo JY**, Baker JA, "A computer aid for diagnosis of breast mass lesions using both mammographic and sonographic BI-RADS descriptors," *Radiology*, (submitted)

3 **Jesneck JL**, Nolte LW, Baker JA, **Lo JY**, "The effect of data set size on computer-aided diagnosis of breast cancer: Comparing decision fusion to a linear discriminant," in SPIE medical Imaging 2006: Image Processing (2006)

REFRENCES

1. S. Shapiro, "Screening: assessment of current studies", *Cancer* 1994; 74:231–238.
2. I.C. Henderson, "Breast cancer", In: Murphy GP, W. Lawrence Jr, R.E. Lenhard, eds. *American Cancer Society textbook of clinical oncology*. Atlanta, Ga: American Cancer Society, 1995; 198–219.
3. A.M. Knutzen, J.J Gisvold, "Likelihood of malignant disease for various categories of mammographically detected, nonpalpable breast lesions," *Mayo Clin Proc* 1993; 68: 454-460.

4. D.B. Kopans, “The Positive Predictive Value of Mammography,” *AJR* 1992; 158:521-526.
5. A. S. Hong, E. L. Rosen, M. S. Soo et al., “BI-RADS for Sonography: Positive and Negative Predictive Values of Sonographic Features.” *AJR* 184 (4), 1260 (2005).
6. E.F. Petricoin, C.P. Paweletz, and L.A. Liotta, “Clinical Applications of Proteomics: Proteomic Pattern Diagnostics”, *Journal of Mammary Gland Biology and Neoplasia*, Vol. 7, No. 4, October 2002, p.433-440.
7. J.Y. Lo, M. Gavrielides, M.K. Markey, J.L. Jesneck, “Computer-aided classification of breast microcalcification clusters: merging of features from image processing and radiologists”, *Proc. SPIE* Vol. 5032, p. 882-889, Medical Imaging 2003: Image Processing; Milan Sonka, J. Michael Fitzpatrick; Eds.
8. BI-RADS. American College of Radiology. *American College of Radiology Breast Imaging - Reporting and Data System (BI-RADS) 3rd ed.*, 1998.

APPENDICES

Three publications are attached, see "Reportable Outcomes" above for the list.

Appendix

Jesneck JL, Nolte LW, Baker JA, Floyd CE, Lo JY, “An optimized approach to decision fusion of heterogeneous data for Breast Cancer Diagnosis,” *Medical Physics*, (in press)…… **10**

Jesneck JL, Lo JY, Baker JA, “A computer aid for diagnosis of breast mass lesions using both mammographic and sonographic BI-RADS descriptors,” *Radiology*, (submitted)..... **50**

Jesneck JL, Nolte LW, Baker JA, Lo JY, “The effect of data set size on computer-aided diagnosis of breast cancer: Comparing decision fusion to a linear discriminant,” in SPIE medical Imaging 2006: Image Processing (2006)…….. **83**

An Optimized Approach to Decision Fusion of Heterogeneous Data for Breast Cancer Diagnosis

Jonathan L. Jesneck, B.S.E.[1,2], Loren W. Nolte, Ph.D.[1,3], Jay A. Baker, M.D.[2], Carey E. Floyd, Jr., Ph.D.[1,2,4], Joseph Y. Lo, Ph.D.[1,2,4]

1. Department of Biomedical Engineering
2. Duke Advanced Imaging Labs, Department of Radiology
3. Department of Electrical and Computer Engineering
4. Medical Physics Graduate Program

March 2006

Running Title: An Optimized Approach to Decision Fusion of Heterogeneous Data

Abstract

As more diagnostic testing options become available to physicians, it becomes more difficult to combine various types of medical information together in order to optimize the overall diagnosis. To improve diagnostic performance, here we introduce an approach to optimize a decision-fusion technique to combine heterogeneous information, such as from different modalities, feature categories, or institutions. For classifier comparison we used two performance metrics: the ROC area under the curve (AUC) and the normalized partial area under the curve (pAUC). This study used four classifiers: linear discriminant analysis (LDA), artificial neural network (ANN), and two variants of our decision-fusion technique, AUC-optimized (DF-A) and pAUC-optimized (DF-P) decision fusion. We applied each of these classifiers with 100-fold cross validation to two heterogeneous breast cancer data sets: one of mass lesion features and a much more challenging one of microcalcification lesion features. For the calcification data set, DF-A outperformed the other classifiers in terms of AUC ($p < 0.02$) and achieved AUC = 0.85 ± 0.01. The DF-P surpassed the other classifiers in terms of pAUC ($p < 0.01$) and reached pAUC = 0.38 ± 0.02. For the mass data set, DF-A outperformed both the ANN and the LDA ($p < 0.04$) and achieved AUC = 0.94 ± 0.01. Although for this data set there were no statistically significant differences among the classifiers' pAUC values (pAUC = 0.57 ± 0.07 to 0.67 ± 0.05, $p > 0.10$), the DF-P did significantly improve specificity versus the LDA at both 98% and 100% sensitivity ($p < 0.04$). In conclusion, decision fusion directly optimized clinically significant performance measures such as AUC and pAUC, and sometimes outperformed two well known machine-learning techniques when applied to two different breast cancer data sets.

Keywords

Decision Fusion, Heterogeneous Data, Receiver Operating Characteristic (ROC) Curve, Area Under the Curve (AUC), Partial Area Under the Curve (pAUC), Classification, Machine Learning, Breast Cancer

I. Introduction

Breast cancer accounts for one-third of all cancer diagnoses among American women, has the second highest mortality rate of all cancer deaths in women [1], and is expected to account for 15% of all cancer deaths in 2005 [2]. Early diagnosis and treatment can significantly improve the chance of survival for breast cancer patients [3]. Currently, mammography is the preferred screening method for breast cancer. However, high false positive rates reduce the effectiveness of screening mammography, as several studies have shown that only 13-29% of suspicious masses are determined to be malignant [4 5 6]. Unnecessary surgical biopsies are expensive, cause patient anxiety, alter cosmetic appearance, and can distort future mammograms [7].

Commercial products for computer-aided detection (CAD) have shown promise for improving sensitivity in large clinical trials. Most studies to date have shown CAD to boost radiologists' lesion detection sensitivity [8 9 10 11]. To date, however, there are no commercial systems to improve specificity for breast cancer screening. To fill this need to improve the sensitivity of mammography, computer-aided diagnosis (CADx) has emerged as a promising clinical aid [12].

There has been considerable CAD and CADx research based upon a rich variety of modalities and sources of medical information such as: digitized screen-film

mammograms [13 14 15 16 17], full-field digital mammograms [18], sonograms [19 20 21], MRI images [22], and gene expression profiles [23]. Current clinically implemented CADx programs tend to use only one information source, although multimodality CADx programs [24] are beginning to emerge. Moreover, most CADx research has been performed using relatively homogeneous data sets collected at one institution, acquired using one type of digitizer or digital detector, or using features drawn from one source such as human-interpreted findings versus computer-extracted features. Increasingly however there is a trend towards boosting diagnostic performance by combining together data from many different sources to create heterogeneous data. We defined heterogeneous data as comprising multiple, distinct groups. Specifically, for this study we considered as heterogeneous any of the following data set characteristics: multiple imaging modalities, multiple types of mammogram film digitizers, data collected from multiple institutions, and various types of features extracted from the same image, especially computer-extracted and human-extracted features. Combining heterogeneous data types for classification is a difficult machine-learning problem, but one that has shown promise in bioinformatics applications [25 26 27].

To meet the challenge of combining heterogeneous data types, we turned to a decision-fusion method that operates by the following two steps: 1. Classifiers use feature subsets to generate initial binary decisions, and 2. These binary decisions are then combined optimally using decision fusion theory. Decision fusion offers the following advantages: It handles heterogeneous data sources well, reduces the problem dimensionality, is easily interpretable, and is easy to use in a clinical setting. Decision fusion has effectively combined heterogeneous data in many diverse classification tasks,

such as detecting land mines using multiple sensors [28], identifying persons using multiple biometrics [29], and CADx of endoscopic images using multiple sets of medical features [30].

The purpose of this study was to optimize a decision-fusion approach for classifying heterogeneous breast cancer data. We compared this decision-fusion approach to a linear discriminant and an artificial neural network, which are well-studied techniques that have frequently been applied to breast cancer CADx [13 31 32 33]. This study evaluates these classification algorithms on two breast cancer data sets using two different clinically relevant performance metrics.

II. Methods

A. Data

For this study, we chose two different breast cancer data sets, which differed considerably in the type and number of patient cases as well as the type and number of medical information features describing those cases.

Microcalcification Lesions

Data set C consisted of all 1508 mammogram microcalcification lesions from the Digital Database for Screening Mammography (DDSM) [34]. The outcomes were verified by histological diagnosis and follow-up for certain benign cases, yielding 811 benign and 697 malignant calcification lesions. Figure 1 shows the feature group structure of this data set. The feature groups were 13 computer-extracted calcification cluster morphological features, 91 computer-extracted texture features of the lesion background anatomy, 2 radiologist-interpreted findings, 3 radiologist-extracted features from the Breast Imaging Reporting and Data System (BI-RADS™, American College of Radiology,

Reston, VA) [35] and patient age. In total, data set C had 110 features and a sample-to-feature ratio of approximately 14:1. Each mammogram was digitized with one of four digitizers: a DBA M2100 ImageClear at a resolution of 42 microns, a Howtek 960 at 43.5 microns, a Howtek MultiRad850 at 43.5 microns, or a Lumisys 200 Laser at 50 microns. To study this large, heterogeneous data set, no attempt was made to restrict cases only to a single digitizer, as was common in most previous studies. Moreover, no standardization step was applied to the images to correct for the differences in noise, resolution, and other physical characteristics from the various digitizers. We used a 512x512 pixel ROI centered on the centroid of each lesion (using lesion outlines drawn by the DDSM radiologists) for image processing and for generating the computer-extracted features. We extracted morphological and texture (spatial gray level dependence matrix) features, which were shown to be useful in a previous study of CADx by Chan et al [31].

This data set had many heterogenic characteristics, such as that it was collected at four different institutions, scanned on four types of digitizers with different physical characteristics, and included both human-extracted and computer-extracted features, such as shape and texture features.

Mass Lesions

Data set M consisted of 568 breast mass cases that were collected in the Radiology Department of Duke University Health System between 1999 and 2001. These cases were an extension of the data set described in detail in our previous studies [36 37]. Definitive histopathologic diagnosis from biopsy was used to determine outcome,

yielding 370 benign and 198 malignant mass lesions. Figure 2 shows the feature group structure of this data set. Dedicated breast radiologists recorded all features.

The mass data set was heterogeneous because it was comprised of 3 distinct types of data: 13 mammogram features, 23 sonogram features in turn drawn from 3 different lexicons (Ultrasound BI-RADS, Stavros, and others) [36], as well as 3 patient history features. In total data set M had 39 features and a sample-to-feature ratio of approximately 15:1.

B. Decision Fusion

There is a growing literature in the area of distributed detection. Although there is even some earlier work, several of the early classical references include the work of Tenney and Sandell, who introduced distributed detection using a fixed fusion processor and optimized the local processors [38]. Chair and Varshney fixed the local processors, and optimized the fusion processor [39]. Reibman and Nolte extended these previous studies by simultaneous optimization of the local detectors while deriving the overall optimum fusion design [40]. Dasarathy summarizes some of the earlier work [41].

Decision fusion theory describes how to combine local binary decisions optimally to determine the presence or absence of a signal in noise [38] [39] [40] [41] [42]. The local binary decisions can come from any arbitrary source.

Figure 3 provides a schematic of our decision-fusion method. Our algorithm is a two-stage process, each with a likelihood ratio calculation. The first stage applies a separate likelihood ratio to each feature. These feature-level likelihood ratios are then compared

to separate thresholds to generate feature-level decisions. These feature-level decisions are then fused in the second stage by computing the likelihood ratio of the binary decision values. The second stage combines the feature-level decisions into one fused likelihood-ratio value, which can be used as a classification decision variable.

Our technique offers the important advantage that it can reduce the dimensionality of the feature space of the classification problem by assigning a classifier to each feature separately. Considering only one feature at a time greatly reduces the complexity of the problem by avoiding the need to estimate multidimensional probability density functions (PDFs) of the feature space. Accurately estimating such multidimensional PDFs likely requires many more observations than a typical medical data set contains. Other benefits of decision fusion are that it is robust in noisy data [43], is not overly sensitive to the likelihood ratio threshold values,[42] and can handle missing data values [44]. Our decision-fusion technique can also be tuned to maximize arbitrary performance metrics (as described later in Section II C) that may be more clinically relevant, unlike more traditional classification algorithms that minimize mean squared error.

1. Detection Theory Approach - the Likelihood Ratio

Although decision fusion combines binary decisions regardless of how those decisions were made, it is still important to choose the right initial classifiers in order to pass as much information to the decision fuser as possible. In our algorithm, we used the likelihood ratio as the initial classifier and applied a threshold to generate the binary decisions on each feature. Previous work has shown the likelihood ratio to be an excellent classifier for breast cancer mass lesion data [45] [46].

According to decision theory, the likelihood ratio is the optimal detector to determine the presence or absence of a signal in noise [47]. For this study, the signal to be detected was the potential malignancy of a breast lesion. The null hypothesis (H_0) was that the signal (malignancy) is not present in the noisy features, while the alternative hypothesis (H_1) was that the signal is present.

$$\begin{aligned} H_0 &: X = N \\ H_1 &: X = S + N \end{aligned} \tag{1}$$

Sources of noise in the features included anatomical noise inherent in the mammogram or sonogram, quantum noise in the acquisition of the mammogram or sonogram, digitization noise and artifacts for data set C, and ambiguities in the mammogram reading process for the radiologist-interpreted findings in both data sets C and M.

The likelihood ratio is the probability of the features under the malignant case divided by the probability of the features under the benign case:

$$\lambda_{features}(X) = \frac{P(X \mid H_1)}{P(X \mid H_0)}, \tag{2}$$

where $P(X \mid H_1)$ is the PDF of the observation data X given that the signal is present, and $P(X \mid H_0)$ is the PDF of the data X given that the signal is not present. The likelihood ratio is optimal under the assumption that the PDFs accurately reflect the true densities. We estimated the one-dimensional PDFs of the features with histograms. We used Scott's rule to determine the optimal histogram bin width, [45]

$$h = 3.5\sigma n^{-1/3}, \tag{3}$$

where h is the bin width, σ is the standard deviation and n is the number of observations. The interval of two standard deviations around the mean, $[\mu - 2\sigma, \mu + 2\sigma]$, was then subdivided by the bin width, h. We assigned the values falling outside this

interval to the extreme left or right bins. Next, we applied a threshold value, τ, to the likelihood ratio to produce a binary decision about the presence of the signal.

$$u = \begin{cases} 1 & \text{if } \lambda_{feature} \geq \tau \\ 0 & \text{if } \lambda_{feature} < \tau \end{cases} \tag{4}$$

2. Fusing the Binary Decisions

For the signal-plus-noise hypothesis H_1, the probability of detecting an existing signal is $P(u=1 \mid H_1) = Pd$ and of missing it is $P(u=0 \mid H_1) = 1 - Pd$. For the noise-only hypothesis H_0, the probability of false detection is $P(u=1 \mid H_0) = Pf$ and of correctly rejecting the missing signal is $P(u=0 \mid H_0) = 1 - Pf$. Using these probabilities, the likelihood ratio value of a binary decision variable has a simple form, as shown in Equation (5).

$$\lambda_{decision}(u) = \frac{P(u \mid H_1)}{P(u \mid H_0)} = \begin{cases} \dfrac{Pd}{Pf} & \text{if } u = 1 \\ \dfrac{1-Pd}{1-Pf} & \text{if } u = 0 \end{cases} \tag{5}$$

We can then use the likelihood ratios of the individual local decision variables to calculate the joint likelihood ratio of the set of decision variables. Assuming that the local decision variables are statistically independent, the likelihood ratio of the fused classifier is a product of the likelihood ratios of the individual local decisions.

$$\lambda_{fusion}(u_1,\ldots,u_p) = \prod_{i=1}^{p} \lambda_{decision}(u_i) = \prod_{i=1}^{p} \frac{P(u_i \mid H_1)}{P(u_i \mid H_0)} = \prod_{i=1}^{p} \left(\frac{Pd_i}{Pf_i}\right)^{u_i} \left(\frac{1-Pd_i}{1-Pf_i}\right)^{1-u_i} \tag{6}$$

Note that we assume statistical independence of only the local binary decisions, not of the sensitivity, false-positive rate, or even the features on which the local decisions were made.

In our decision fusion theory approach we have made the important assumption that all the local decisions are statistically independent. While this appears to be a very strong assumption, using it in decision fusion often does not lower classification performance substantially below the performance of the optimal decision fusion processor for correlated decisions. Although we can construct an optimal correlated decision fusion processor with known decision correlations [48], it is difficult to estimate the correlation structure of the decisions accurately, especially given many decisions but only few observations. However, even with correlated decisions, the simplifying assumption of independent decisions often does not lower decision fusion performance. Liao *et al.* have shown that, under certain conditions for the case of fusing two correlated decisions, the independent fusion processor exactly matched the performance of the optimal correlated decision fusion processor. Even in many situations when the optimality conditions were not kept, the degradation of the fusion performance was not significant [42]. Another benefit of the independent local decisions assumption is that decision fusion can usually recover from weak signals and correlated features given enough decisions to fuse [43]. Because we have a large number of local decisions by setting a separate local decision for each feature, our algorithm takes advantage of this performance benefit.

C. Classifier Evaluation and Figures of Merit

We used the ROC curve to capture the classification performance of our decision-fusion algorithm. Assuming independent local decisions, the probability density functions (PDFs) of the decision fusion likelihood ratio have a similar product form [42]:

$$P(\lambda_{fusion} \mid H_1) = \prod_{i=1}^{p} (Pd_i)^{u_i} (1 - Pd_i)^{1-u_i}$$
$$P(\lambda_{fusion} \mid H_0) = \prod_{i=1}^{p} (Pf_i)^{u_i} (1 - Pf_i)^{1-u_i} \quad (7)$$

Using the fusion likelihood ratio value as a classification decision variable, the probabilities of detection and false alarm are calculated as follows:

$$Pd_{fusion}(\beta) = \sum_{\lambda_{fusion} \geq \beta} P(\lambda = \lambda_{fusion} \mid H_1)$$
$$Pf_{fusion}(\beta) = \sum_{\lambda_{fusion} \geq \beta} P(\lambda = \lambda_{fusion} \mid H_0) \quad (8)$$

where β is a threshold on λ_{fusion} that determines the operating point on the ROC curve. By varying the value of the threshold β, these $Pd_{fusion}(\beta)$ and $Pf_{fusion}(\beta)$ values trace the entire decision-fusion ROC curve.

One can use the ROC curve to quantify classification performance by calculating summary metrics of the curve. Certain performance metrics have more significance in a clinical setting than others, especially when high sensitivity must be maintained. This study used two clinically interesting summary metrics of the ROC curve: the area under the curve (AUC), and the normalized partial area under the curve (pAUC) above a certain sensitivity value [49]. For this study, we set the sensitivity value TPF = 0.90 for pAUC to reflect that diagnosing breast cancer at high sensitivities is clinically imperative. We used the non-parametric bootstrap method [50] to measure the means and variances of the AUC and pAUC values as well as to compare metrics from two models for statistical significance.

D. Genetic Algorithm Search for the Optimal Threshold Set

The selection of the likelihood-ratio threshold values is important to maximize performance of the fused classifier. Threshold values very far from the best values often lowered the fused classifier's performance to near chance levels. A genetic algorithm searched over the likelihood-ratio threshold values for each feature to select a threshold set that maximized the desired performance metric or figure of merit (FOM),

$$\tau_{optimal} = \text{argmax FOM}(\lambda_{fusion}(u;\tau)), \tag{8}$$

where the FOM is either AUC or pAUC, u is the set of local decisions, and τ is the set of feature-level likelihood-ratio thresholds. The fitness function of the genetic algorithm was set to the FOM in order to maximize the FOM value. We optimized for cross-validation performance the following genetic algorithm parameters: the number of generations, population size, and rates of selection, crossover, and mutation.

E. Decision Fusion with Cross Validation

We used k-fold cross validation (k=100) to estimate the ability of the classifiers to generalize on our data sets. For each fold, a new model was developed, i.e., the likelihood ratio was formed on the k-1 subsets (99% of cases) used as training samples, and the genetic algorithm searched over the thresholds to maximize the performance metric on these training samples. Once the best thresholds had been found on the training set, they were then used to evaluate the algorithm on the one subset (1% of cases) withheld for validation. The resulting local decisions were then combined into the fused validation likelihood ratio $\lambda_{test,fusion}$, as in Equation (6). The process was then repeated k times by withholding a different subset for validation, such that all cases are used for training and validation while simultaneously ensuring independence between those subsets.

Compiling all $\lambda_{test,fusion}$ values at the end of the cross validation computations created a distribution of $\lambda_{test,fusion}(X)$ of the test cases. We constructed an ROC curve from the $\lambda_{test,fusion}(X)$ values, as in Equation (8), in order to measure the classification performance of the decision-fusion classifier with k-fold cross validation.

F. Using Decision Fusion in a Diagnostic Setting

Once the model has been fully trained and validated, it can similarly be applied to new cases by setting all of the existing data to be the training data and applying the new clinical case as a new validation case. The decision-fusion algorithm would recommend to the physician either a biopsy with a malignant classification or short-term follow-up with a very likely benign classification.

G. Other Classifiers: Artificial Neural Network and Linear Discriminant

We compared the classification performance of the decision fusion against both an artificial neural network (ANN) and Fisher's linear discriminant analysis (LDA), which are well-understood algorithms and are popular breast cancer CADx research tools.

For the ANN, we used a fully-connected, feed-forward, error backpropagation network with a hidden layer of 5 nodes, implemented using the nnet package (version 7.2-20) for R statistical software (version 1.12, the R Project for Statistical Computing). For the LDA, we used the Statistics Toolbox (version 5.1) of MATLAB® (Release 14, Service Pack 2, Mathworks Inc, Natick MA). Both models were carefully verified against custom software previously developed within our group. We implemented our decision-fusion algorithm in

MATLAB, relying specifically on the Genetic Algorithm and Direct Search Toolbox (version 2) to find the best thresholds for the likelihood ratio values.

III. Results

A. Classifier Performance on Data Set C (Calcification Lesions)

Figure 4 shows the validation ROC curves for the calcification data. Table 1 lists the classification performances of the four classifiers, while Tables 2 and 3 list the two-tailed p-values for the pairwise comparisons by AUC and pAUC, respectively. The DF-A showed the best overall performance, with AUC = 0.85 ± 0.01, and the DF-P was slightly worse with AUC = 0.82 ± 0.01. Both decision-fusion ROC curves were well above those of the LDA and ANN, both in terms of AUC ($p < 0.0001$) and pAUC ($p < 0.02$). None of the features were particularly strong by themselves; we ran an LDA on each feature separately, yielding on average AUC = 0.53 ± 0.03, with a maximum of AUC = 0.66 for the best feature.

The DF-P curve (pAUC = 0.38 ± 0.02) crossed the DF-A curve (pAUC = 0.28 ± 0.03) at the line TPF = 0.9. In order to gain high-sensitivity performance, DF-P sacrificed performance in the less clinically relevant range of TPF < 0.9. The DF-A beat the DF-P in terms of AUC ($p = 0.018$) but lost in pAUC ($p < 0.01$). Both decision-fusion classifiers greatly outperformed the both the ANN (pAUC = 0.14 ± 0.02) and LDA (pAUC = 0.09 ± .06) in terms of pAUC.

B. Classifier Performance on Data Set M (Mass Lesions)

Figure 5 shows the validation ROC curves of the classifiers for the mass data set. Table 4 lists the classification performances of the four classifiers, whereas Tables 5 and 6 list the p-values for the pairwise comparisons by AUC and pAUC, respectively. For this data set, all the classifiers had higher but very similar performance, with AUC ranging from 0.93 ± 0.01 (LDA) to 0.94 ± 0.01 (DF-A). With the exception of DF-P ($p = 0.50$), the DF-A nonetheless significantly outperformed both the LDA ($p = 0.021$) and the ANN ($p = 0.038$) in terms of AUC. The LDA, ANN, and DF-P curves were all very similar, for both AUC ($p > 0.10$) and pAUC ($p > 0.10$). Figure 5 (b) shows the ROC curves in the high sensitivity region above the line TPF = 0.90. The classifiers' pAUC values ranged narrowly from 0.57 ± 0.07 (ANN) to 0.67 ± 0.05 (DF-P), all close enough to show no statistically significant differences ($p > 0.10$). However, the DF-P did have a higher specificity than the LDA at both 98% sensitivity (0.37 ± 0.10 vs. 0.13 ± 0.13, $p = 0.04$) and at 100% sensitivity (0.34 ± 0.08 vs. 0.09 ± 0.12, $p = 0.03$). The DF-P curve passed the DF-A curve approximately at the line TPF = 0.90 and yielded a slightly higher pAUC (0.67 ± 0.05 vs. 0.63 ± 0.07), although this improvement was not statistically significant ($p = 0.48$).

IV. Discussion

The multitude of medical data becoming available to physicians presents the problem of how best to integrate the information for diagnostic performance. Despite recent availability of this information, current CADx programs for breast cancer tend to use only one type of data, usually digitized mammogram films. Because many clinical tests provide complementary information about a disease state, it is important to develop a CADx system that incorporates data from disparate sources. However, combining

disparate data types together for classification is a difficult machine-learning problem. This study used the likelihood-ratio detector and decision-fusion classifier to detect the presence of a malignancy (a signal) within medical data (noisy features). We also compared the performance of this classifier to two popular classifiers in the CADx literature, LDA and ANN, and we measured the diagnostic performance with two classification metrics, ROC AUC and pAUC. Finally, we performed these studies using two very different data sets in order to assess performance differences due to the data set itself.

Data set C (calcification lesions) had a stronger nonlinear component, indicated by the fact that the ANN AUC was much greater than the LDA AUC. The robustness of the decision-fusion algorithm is evident in its good performance on this weaker, nonlinear, and noisy data set. Decision fusion significantly outperformed the ANN and LDA on the calcification data set for both performance metrics. Figure 4 and Table 1 show that the biggest performance gain is in the pAUC metric, for which decision fusion doubled the performance of the other classifiers.

On data set M (mass lesions), all four classifiers seemed to be saturated at a high level of performance in terms of both AUC and pAUC, as shown in Figure 5 and Table 4. Performances were largely equivalent across all models, except for two trends. In terms of AUC, the DF-A outperformed both the ANN and the LDA ($p = 0.038$ and 0.021, respectively). Although on this data set decision fusion offered only relatively modest gains in pAUC, it did achieve a significantly better specificity than the LDA at several of the highest sensitivities of the ROC curve ($p < 0.05$).

This decision-fusion algorithm has many potential benefits over more traditional classification algorithms. Decision fusion can be optimized for any desired performance metric by incorporating the metric into the fitness function of the genetic algorithm for its search over the likelihood-ratio thresholds. This advantage has important clinical implications, as both the physician and the CADx algorithm are constrained to operate at high sensitivity. The performance metric can emphasize good performance at high sensitivities and deemphasize performance at clinically unacceptable low sensitivities. Therefore we expect the DF-A curve to maximize AUC and the DF-P curve to maximize pAUC. The DF-P curve should fall under the DF-A curve for low FPF values but should cross the DF-A curve at the line TPF=0.90 to capture a greater pAUC value. Figures 4 and 5 show evidence that the DF-P did optimize pAUC. The DF-P ROC curves crossed the DF-A curves at the line TPF = 0.90 and do in fact have a larger pAUC value than the DF-A curves. Another advantage is that decision fusion is robust and can recover from noisy, weak features. The likelihood-ratio classifier passes information about the strength or weakness of a feature to the decision fuser, which adjusts the influence given to that feature. This feature-strength information is the ROC operating point (sensitivity and specificity) determined by the likelihood-ratio threshold that was found by the genetic algorithm search. Figure 3 shows a schematic of this information flow from the individual features to the decision fuser. The robustness of the algorithm also suggests that decision fusion may be able to reach the asymptotic validation performance with fewer data. This is important for most medical researchers who are starting to collect new databases and for any databases that are expensive to collect. Because our decision-fusion technique needs to estimate only one-dimensional PDFs, which require much fewer data points than multidimensional PDFs, decision fusion needs many fewer data points for training. For this reason, the decision-fusion algorithm may be able to handle

typical clinical data sets with missing data, as shown in previous work with decision fusion [44].

Drawbacks of the decision-fusion algorithm include losing potentially useful feature information by reducing the likelihood-ratio values of the features to a binary value. Although the algorithm loses some feature information in this step, it recovers by optimally fusing the remaining binary feature information from that point forward. In the ideal case, if the true underlying multivariate distribution of the data happens to be known or can be estimated with a high degree of confidence, then the Bayes classifier can take this information into account and is theoretically optimal. However, since the true underlying distribution is almost never known in practice, decision fusion is a good alternative method, especially for small and noisy data sets.

V. Conclusions

We have developed a decision-fusion classification technique that combines features from heterogeneous data sources. We have demonstrated the technique on both a data set of two different breast imaging modalities and a data set of human-extraced versus computer-extracted findings. With our data, decision fusion always performed as well as or better than the classic classification techniques LDA and ANN. The improvements were all significant for the more challenging data set C, but not always significant for the less challenging data set M. Such a statement may not reflect the full diversity of these data sets, which differ in many respects, including linear separability, numbers of cases and features, and feature correlations. Future work will explore the contribution of such factors in order to understand the full potential and limitations of the decision-fusion

technique. In conclusion, the decision-fusion technique showed particular strength in the task of combining groups of weak, noisy features for classification.

Acknowledgments

This work was supported by US Army Breast Cancer Research Program W81XWH-05-1-0292 and DAMD17-02-1-0373, and NIH/NCI R01 CA95061 and R21 CA93461. We thank Brian Harrawood for the ROC bootstrap code, Anna Bilska-Wolak, Ph.D., and Georgia Tourassi, Ph.D., for insightful discussions, and Andrea Hong, M.D., Jennifer Nicholas, M.D., Priscilla Chyn, and Susan Lim for data collection.

Table Legends

Table 1. Classifier Performance on Data Set C (Calcification Lesions)
The table shows the AUC and pAUC values for the ROC curves of the four classifiers under 100-fold cross validation. The performance values exhibited a wide range. The DF-A scored the best for AUC, while DF-P scored highest for pAUC, as expected. The decision fusion curves soundly outperformed both the ANN and LDA in terms of pAUC.

Table 2. P-values for AUC Comparisons for Data Set C (Calcification Lesions)
The confusion matrix shows the p-values for the pairwise comparisons of the classifiers' AUC values. All pairwise comparisons were statistically significant.

Table 3. P-values for pAUC Comparisons for Data Set C (Calcification Lesions)
The confusion matrix shows the p-values for the pairwise comparisons of the classifiers' pAUC values. All pairwise comparisons were statistically significant.

Table 4. Classifier Performance on Data Set M (Mass Lesions)
The table shows the AUC and pAUC values for the ROC curves of the four classifiers under 100-fold cross validation. All four classifiers performed very similarly on this data set. The DF-A scored the best for AUC, whereas the DF-P scored highest for pAUC, although both were still within one standard deviation of each of the other classifiers' performances.

Table 5. P-values for AUC Comparisons for Data Set M (Mass Lesions)

The confusion matrix shows the p-values for the pairwise comparisons of the classifiers' AUC values. The DF-A outperformed the ANN and LDA. Among the DF-P, ANN, and LDA, there were no statistically significant pAUC differences.

Table 6. P-values for pAUC Comparisons for Data Set M (Mass Lesions)

The confusion matrix shows the p-values for the pairwise comparisons of the classifiers' pAUC values. None of the pAUC comparisons were statistically significant. Although pAUC scores were similar, the DF-P did have a higher specificity than the LDA at both 98% sensitivity (0.37 ± 0.10 vs. 0.13 ± 0.13, $p = 0.04$) and at 100% sensitivity (0.34 ± 0.08 vs. 0.09 ± 0.12, $p = 0.03$).

Tables

Table 1. Classifier Performance on Calcification Data Set C

Classifier	AUC	pAUC
DF-A	0.85 ± 0.01	0.28 ± 0.03
DF-P	0.82 ± 0.01	0.38 ± 0.02
ANN	0.76 ± 0.01	0.14 ± 0.02
LDA	0.68 ± 0.01	0.09 ± 0.06

Table 2. P-values for AUC Comparisons for Calcification Data Set C

	DF-A	DF-P	ANN	LDA
DF-A		0.018	< 0.0001	< 0.0001
DF-P			0.0001	< 0.0001
ANN				< 0.0001
LDA				

Table 3. P-values for pAUC Comparisons for Calcification Data Set C

	DF-A	DF-P	ANN	LDA
DF-A		0.0084	0.018	< 0.0001
DF-P			0.0001	< 0.0001
ANN				0.016
LDA				

Table 4. Classifier Performance on Mass Data Set M

Classifier	AUC	pAUC
DF-A	0.94 ± 0.01	0.63 ± 0.07
DF-P	0.93 ± 0.01	0.67 ± 0.05
ANN	0.93 ± 0.01	0.57 ± 0.07
LDA	0.93 ± 0.01	0.59 ± 0.06

Table 5. P-values for AUC Comparisons for Mass Data Set M

	DF-A	DF-P	ANN	LDA
DF-A		0.50	0.038	0.021
DF-P			0.20	0.17
ANN				0.53
LDA				

Table 6. P-values for pAUC Comparisons for Mass Data Set M

	DF-A	DF-P	ANN	LDA
DF-A		0.48	0.45	0.27
DF-P			0.14	0.12
ANN				0.46
LDA				

Figure Legends

Figure 1. Feature Group Structure for Calcification Data Set C (Calcification Lesions)
The features of the calcification data set consisted of three main groups: computer-extracted features, radiologist-extracted features, and patient history features. The computer-extracted features were morphological and shape features of the automatically detected and segmented microcalcification clusters within the digitized mammogram images. The radiologist-extracted features comprised both radiologist-interpreted findings and BI-RADS features. This data set consisted of 512x512 pixel ROIs of all 1508 calcification lesions in the Digital Database for Screening Mammography (DDSM). This data set had many heterogenic characteristics, such as that it was collected at four different institutions, scanned on four digitizers with different noise characteristics, and included both human-extracted and computer-extracted features, such as shape and texture features.

Figure 2. Feature Group Structure for Mass Data Set M (Mass Lesions)
The features of the mass data set consisted of mammogram features, sonogram features, and patient history features. The mammogram features comprised both BI-RADS features and radiologist-interpreted findings. The sonogram features consisted of ultrasound BI-RADS features, Stavros features, and other ultrasound mass descriptors. All image features were radiologist-extracted features. The mass data set was heterogeneous in including both mammogram and sonogram views of the breast. Both mammogram and sonogram feature sets were as well as including patient history features.

Figure 3. The Role of Likelihood-ratio Thresholds for Decision Fusion

The first column shows plots of the log-likelihood-ratio vs. feature value for each feature. The algorithm calculated the likelihood ratio and then thresholded it separately for each feature. The threshold determined the ROC operating point of the likelihood-ratio classifier of a particular feature. Next, the algorithm combined the binary decisions from the feature-level likelihood ratio classifiers using decision fusion theory to produce the likelihood ratio of the fused classifier.

Figure 4. ROC Curves for Data Set C (Calcification Lesions)

The classifiers' ROC curves for 100-fold cross validation are shown. Figure 2 (a) shows the full ROC curves, while Figure 2 (b) shows only the high-sensitivity region (TPF ≥ 0.90). For the calcification data set, the four classifiers yielded differing classification performance under 100-fold cross validation. Both decision-fusion curves lay significantly above the LDA and ANN curves, both in terms of AUC and pAUC. As expected, the decision-fusion classifiers achieved the highest scores of all the classifiers for their target performance metrics; DF-A attained the greatest AUC, whereas DF-P attained the greatest pAUC. The DF-P curve surpassed the DF-A curve and dominated the other curves above the line TPF = 0.90. In order to gain high-sensitivity performance, DF-P sacrificed performance in the less clinically relevant range of TPF < 0.90.

Figure 5. ROC Curves for Data Set M (Mass Lesions)

For the mass data set, all classifiers had high levels of classification performance. The DF-A and DF-P achieved the highest AUC and pAUC, respectively. In terms of AUC, the DF-A outperformed both the ANN and LDA ($p = 0.038$ and 0.021, respectively). In Figure 5 (b), the DF-P curve had slightly more partial area than the other curves. Despite having

statistically equivalent partial areas, the DF-P had a greater specificity than the LDA at high sensitivities TPF = 0.98 ($p = 0.03$).

Figures

Figure 1. Feature Group Structure for Calcification Data Set C (Calcification Lesions)

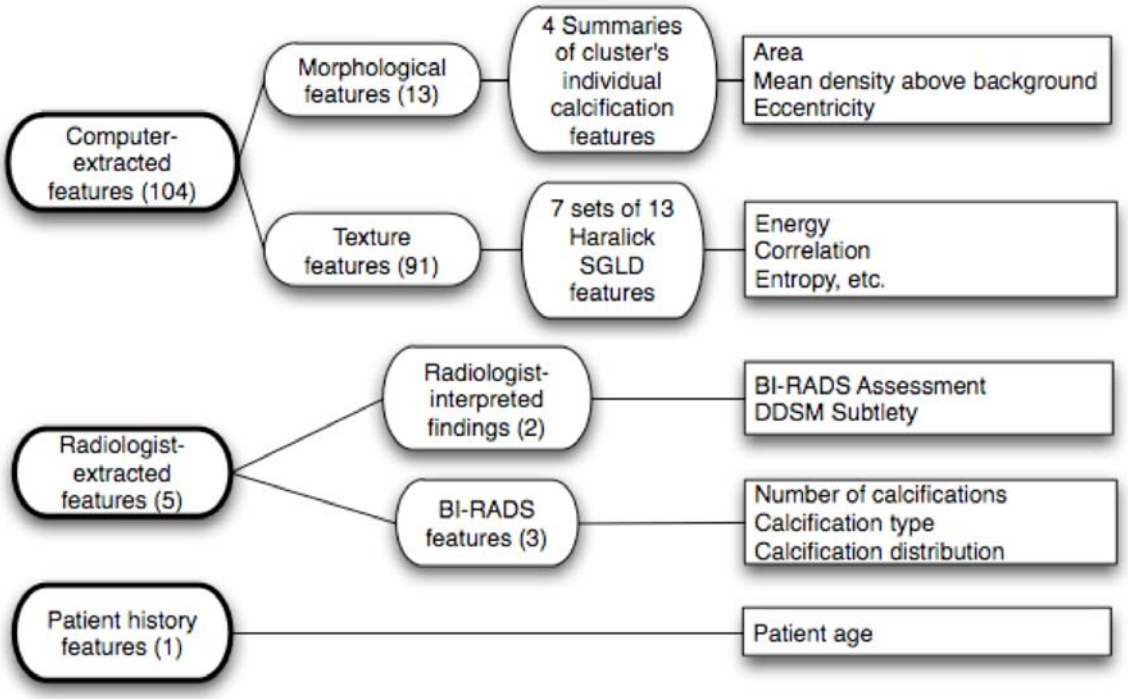


Figure 2. Feature Group Structure for Mass Data Set M (Mass Lesions)

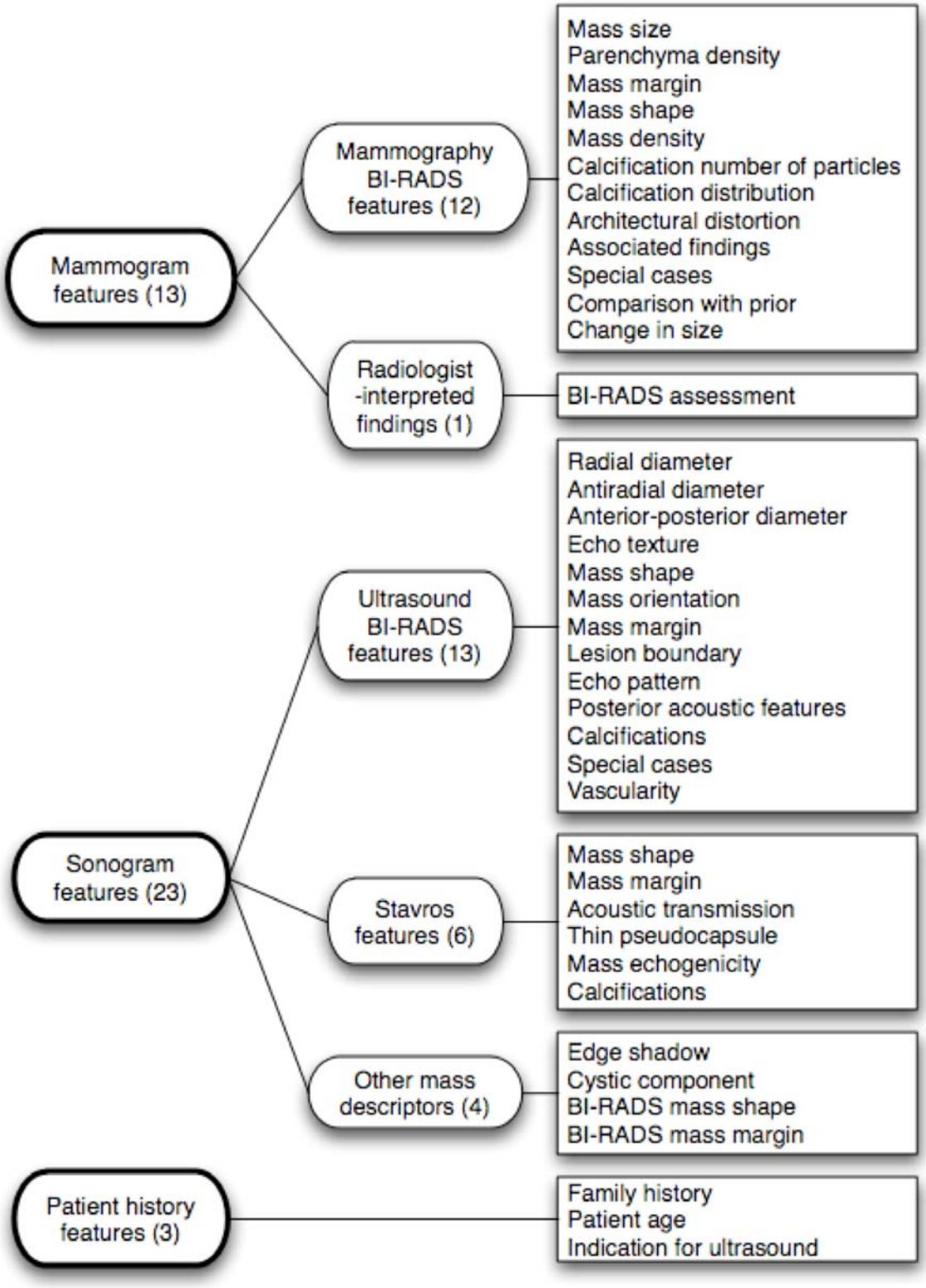


Figure 3. The Role of Likelihood-ratio Thresholds for Decision Fusion

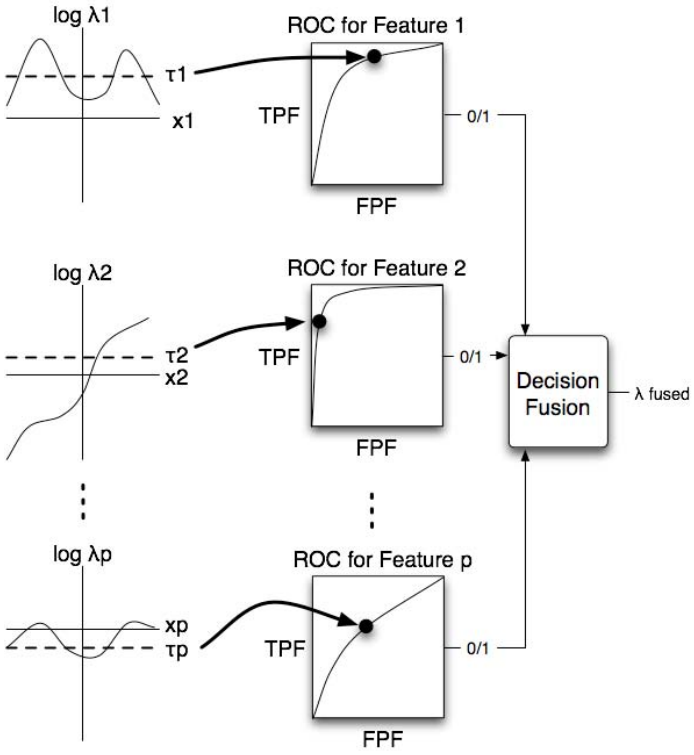
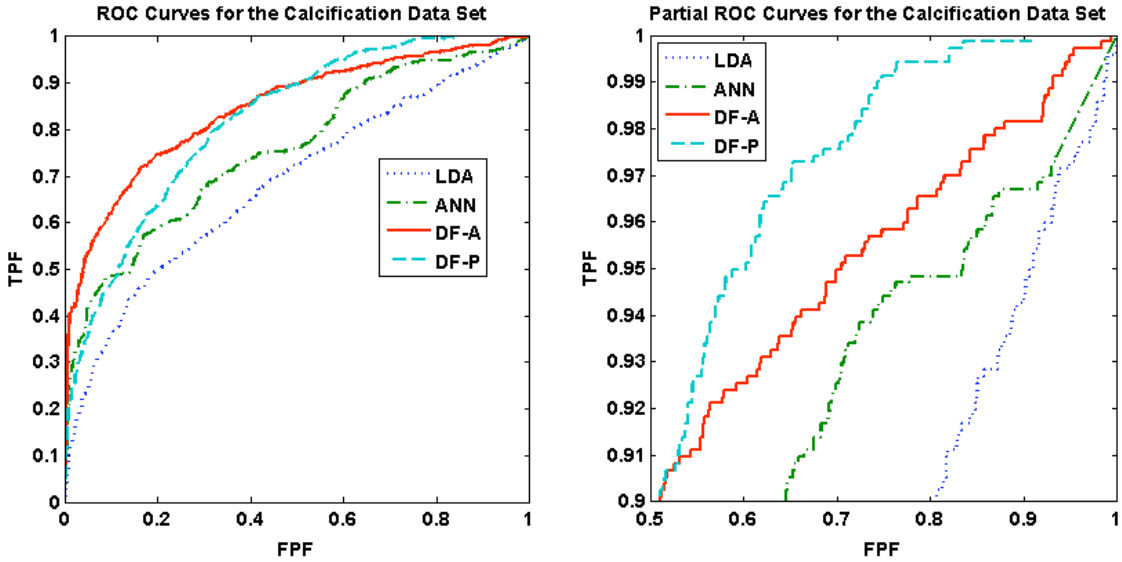
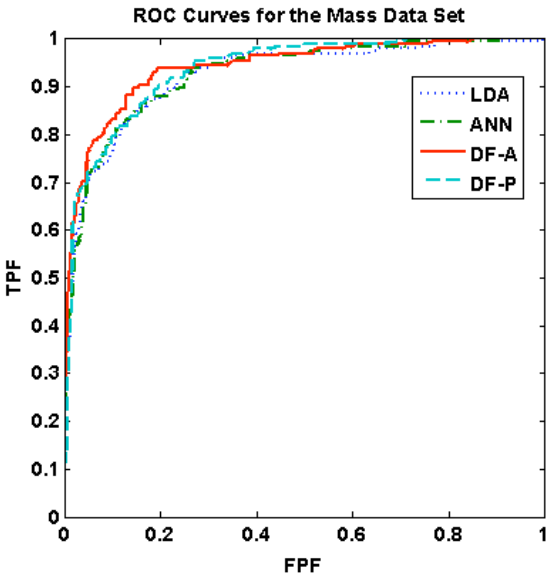


Figure 4. ROC Curves for Data Set C (Calcification Lesions)

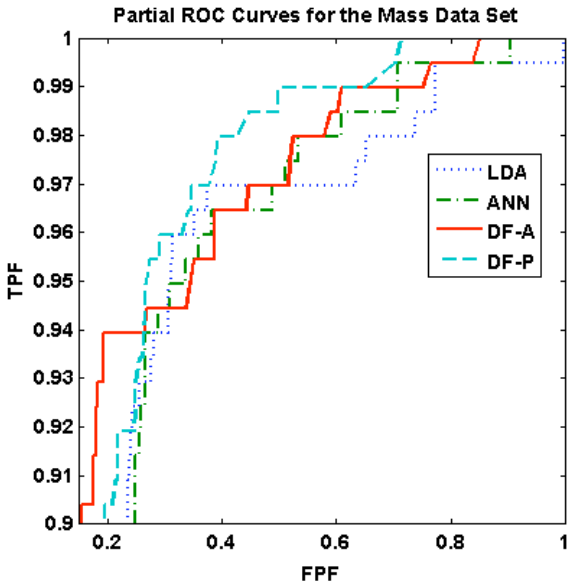


(a) ROC curves (b) Partial ROC curves (TPF ≥ 0.90)

Figure 5. ROC Curves for Data Set M (Mass Lesions)



(a) ROC curves



(b) Partial ROC curves (TPF ≥ 0.90)

Abbreviations

ANN	Artificial Neural Network
AUC	Area Under the ROC curve
CAD	Computer-aided Detection
CADx	Computer-aided Diagnosis
DDSM	Digital Database for Screening Mammography
DF-A	AUC-optimized Decision Fusion
DF-P	pAUC-optimized Decision Fusion
FPF	False Positive Fraction
LDA	Linear Disciminant Analysis
pAUC	Partial Area Under the ROC curve (TPF ≥ 0.90)
Pd	Probability of Detection
Pf	Probability of False Alarm
ROC	Receiver Operating Characteristic
SGLD	Spatial Gray Level Dependence
TPF	True Positive Fraction

VI. References

[1] J.V. Lacey, Jr., S.S. Devesa, and L.A. Brinton, "Recent trends in breast cancer incidence and mortality." *Env. Mol. Mutag.* **39**, 82 (2002).

[2] A. Jemal, T. Murray, E. Ward, A. Samuels, R.C. Tiwari, A. Ghafoor, E.J. Feuer, and M.J. Thun, "Cancer statistics, 2005." *Ca: a Cancer Journal for Clinicians* **55**, 10 (2005).

[3] B. Cady and J.S. Michaelson, "The life-sparing potential of mammographic screening." *Cancer* **91**, 1699 (2001).

[4] J.E. Meyer, D.B. Kopans, P.C. Stomper, and K.K. Lindfors, "Occult breast abnormalities: percutaneous preoperative needle localization." *Radiology* **150**, 335 (1984).

[5] A.L. Rosenberg, G.F. Schwartz, S.A. Feig, and A.S. Patchefsky, "Clinically occult breast lesions: localization and significance." *Radiology* **162**, 167 (1987).

[6] B.C. Yankaskas, M.H. Knelson, J.T. Abernethy, J.T. Cuttino, and R.L. Clark, "Needle localization biopsy of occult lesions of the breast." *Radiology* **23**, 729 (1988).

[7] M.A. Helvie, D.M. Ikeda, and D.D. Adler, "Localization and needle aspiration of breast lesions: complications in 370 cases." *Am. J. Roentgenol.* **157**, 711 (1991).

[8] L.J.W. Burhenne, S.A. Wood, C.J. D'Orsi, S.A. Feig, D.B. Kopans, K.F. O'Shaughnessy, E.A. Sickles, L. Tabar, C.J. Vyborny, and R.A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography." *Radiology* **215**, 554 (2000).

[9] T.W. Freer and M.J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center." *Radiology* **220**, 781 (2001).

[10] R.F. Brem, J. Baum, M. Lechner, S. Kaplan, S. Souders, L.G. Naul, and J. Hoffmeister, "Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial." *Am. J. Roentgenol.* **181**, 687 (2003).

[11] S.V. Destounis, P. DiNitto, W. Logan-Young, E. Bonaccio, M.L. Zuley, and K.M. Willison, "Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience." *Radiology* **232**, 578 (2004).

[12] C.J. Vyborny, "Can computers help radiologists read mammograms?" *Radiology* **191**, 315 (1994).

[13] H.P. Chan, B. Sahiner, N. Petrick, M.A. Helvie, K.L. Lam, D.D. Adler, and M.M. Goodsitt, "Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network." *Phys. Med. Biol.* **42**, 549 (1997).

[14] M.A. Gavrielides, J.Y. Lo, and C.E. Floyd, Jr, "Parameter optimization of a computer-aided diagnosis scheme for the segmentation of microcalcification clusters in mammograms." *Med. Phys.* **29**, 475 (2002).

[15] N. Petrick, H.P. Chan, D. Wei, B. Sahiner, M.A. Helvie, and D.D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and texture classification." *Med. Phys.* **23**, 1685 (1996).

[16] N. Petrick, B. Sahiner, H.P. Chan, M.A. Helvie, S. Paquerault, and L.M. Hadjiiski, "Breast cancer detection: Evaluation of a mass-detection algorithm for computer-aided diagnosis - Experience in 263 patients." *Radiology* **224**, 217 (2002).

[17] Y.H. Chang, B. Zheng, and D. Gur, "Computerized identification of suspicious regions for masses in digitized mammograms." *Invest Radiol* **31**, 146 (1996).

[18] J. Wei, B. Sahiner, L.M. Hadjiiski, H.-P. Chan, N. Petrick, M.A. Helvie, M.A. Roubidoux, J. Ge, and C. Zhou, "Computer-aided detection of breast masses on full field digital mammograms." *Med. Phys.* **32**, 2827 (2005).

[19] D. Chen, R.F. Chang, and Y.L. Huang, "Breast cancer diagnosis using self-organizing map for sonography." *Ultrasound Med. Biol.* **26**, 405 (2000).
[20] K. Horsch, M.L. Giger, L.A. Venta, and C.J. Vyborny, "Computerized diagnosis of breast lesions on ultrasound." *Med. Phys.* **29**, 157 (2002).
[21] K. Horsch, M.L. Giger, C.J. Vyborny, and L.A. Venta, "Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography." *Acad Radiol* **11**, 272 (2004).
[22] W. Chen, M.L. Giger, L. Lan, and U. Bick, "Computerized interpretation of breast MRI: investigation of enhancement-variance dynamics." *Med Phys* **31**, 1076 (2004).
[23] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson, J.R. Marks, and J.R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles." *PNAS* **98**, 11462–11467 (2001).
[24] B. Sahiner, H.-P. Chan, L.M. Hadjiiski, M.A. Roubidoux, C. Paramagul, M.A. Helvie, and C. Zhou, "Multimodality CAD: combination of computerized classification techniques based on mammograms and 3D ultrasound volumes for improved accuracy in breast mass characterization," Medical Imaging 2004: Image Processing, San Diego, CA, USA, **5370**, 67 (2004).
[25] P. Pavlidis, J. Weston, J. Cai, and W.S. Noble, "Learning gene functional classifications from multiple data types." *J. Comp. Biol.* **9**, 401 (2002).
[26] G.R. Lanckriet, T. De Bie, N. Cristianini, M.I. Jordan, and W.S. Noble, "A statistical framework for genomic data fusion." *Bioinformatics* **20**, 2626 (2004).
[27] G.R. Lanckriet, M. Deng, N. Cristianini, M.I. Jordan, and W.S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast." *Pacific Symposium on Biocomputing*, 300 (2004).
[28] Y. Liao, L.W. Nolte, and L. Collins, "Optimal Multisensor Decision Fusion of Mine Detection Algorithms," SPIE 17th Annual AeroSense Symposium, Orlando, FL, United States, **5089**, 1252 (2003).
[29] K. Veeramachaneni, L.A. Osadciw, and P.K. Varshney, "An adaptive multimodal biometric management algorithm." *IEEE Trans. Sys., Man and Cyber. Part C: Applications and Reviews* **35**, 344 (2005).
[30] M.M. Zheng, S.M. Krishnan, and M.P. Tjoa, "A fusion-based clinical decision support for disease diagnosis from endoscopic images." *Comp. Biol. Med.* **35**, 259 (2005).
[31] H.P. Chan, B. Sahiner, K.L. Lam, N. Petrick, M.A. Helvie, M.M. Goodsitt, and D.D. Adler, "Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces." *Med. Phys.* **25**, 2007 (1998).
[32] M. Kallergi, "Computer-aided diagnosis of mammographic microcalcification clusters." *Med. Phys.* **31**, 314 (2004).
[33] M.K. Markey, J.Y. Lo, and C.E. Floyd, Jr, "Differences between computer-aided diagnosis of breast masses and that of calcifications." *Radiology* **223**, 489 (2002).
[34] M. Heath, K.W. Bowyer, and D. Kopans, "Current status of the Digital Database for Screening Mammography," in *Digital Mammography*, edited by N. Karssemeijer, M. Thijssen, and J. Hendriks (Kluwer Academic Publishers, 1998), pp. 457.
[35] BI-RADS, *American College of Radiology Breast Imaging - Reporting and Data System (BI-RADS) 3rd ed.*, American College of Radiology, 1998.
[36] A.S. Hong, E.L. Rosen, M.S. Soo, and J.A. Baker, "BI-RADS for Sonography: Positive and Negative Predictive Values of Sonographic Features." *Am. J. Roentgenol.* **184**, 1260 (2005).
[37] J.L. Jesneck, J.Y. Lo, and J.A. Baker, "A computer aid for diagnosis of breast mass lesions using both mammographic and sonographic BI-RADS descriptors." *Radiology*, submitted (2006).
[38] R.R. Tenney and N.R. Sandell, Jr., *Detection with Distributed Sensors*. (IEEE, Piscataway, NJ, Albuquerque, NM, 1980).

[39] Z. Chair and P.K. Varshney, "Optimal data fusion in multiple sensor detecton systems." *IEEE Trans. Aero. Elec. Sys.* **AES-22**, 98 (1986).
[40] A.R. Reibman and L.W. Nolte, "Optimal detection and performance of distributed sensor systems." *IEEE Trans. Aero. Elec. Sys.* **AES-23**, 24 (1987).
[41] B.V. Dasarathy, "Decision fusion strategies in multisensor environments." *IEEE Trans. Sys., Man and Cyber.* **21**, 1140 (1991).
[42] Y. Liao, *Distributed decision fusion in signal detection -- a robust approach*, Ph.D. Thesis, Duke University, 2005.
[43] R. Niu, P.K. Varshney, M. Moore, and D. Klamer, "Decision fusion in a wireless sensor network with a large number of sensors," Proceedings of the Seventh International Conference on Information Fusion, FUSION 2004, Stockholm, Sweden, **1**, 21 (2004).
[44] A.O. Bilska-Wolak and C.E. Floyd, Jr., "Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer." *Phys Med Biol* **49**, 4219 (2004).
[45] A.O. Bilska-Wolak, C.E. Floyd, Jr, L.W. Nolte, and J.Y. Lo, "Application of likelihood ratio to classification of mammographic masses; performance comparison to case-based reasoning." *Med. Phys.* **30**, 949 (2003).
[46] A.O. Bilska-Wolak, C.E. Floyd, Jr., J.Y. Lo, and J.A. Baker, "Computer aid for decision to biopsy breast masses on mammography: validation on new cases." *Acad Radiol* **12**, 671 (2005).
[47] H.L. VanTrees, *Detection, Estimation, and Modulation Theory (Part I)*. (John Wiley & Sons, New York, 1968).
[48] E. Drakopoulos and C.-C. Lee, "Optimum multisensor fusion of correlated local decisions." *IEEE Trans. Aero. Elec. Sys.* **27**, 593 (1991).
[49] Y. Jiang, C.E. Metz, and R.M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests." *Radiology* **201**, 745 (1996).
[50] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*. (Chapman & Hall, New York, NY, 1993).

A Computer Aid for Diagnosis of Breast Mass Lesions Using Both Mammographic and Sonographic BI-RADS Descriptors

Jonathan L. Jesneck[1,2], Joseph Y. Lo, Ph.D.[1,2], Jay A. Baker, M.D.[2]

1. Department of Biomedical Engineering, Duke University
2. Duke Advanced Imaging Labs, Department of Radiology, Duke University Health System

April 2006

Duke Advanced Imaging Labs
Duke University Medical Center
2424 Erwin Road, Suite 302
Durham, NC 27705

Tel. 919-684-1440
Fax. 919-684-1491

Email: jonathan.jesneck@duke.edu

This work was supported by US Army Breast Cancer Research Program W81XWH-05-1-0292 and DAMD17-02-1-0373, and NIH/NCI R01 CA95061 and R21 CA93461.

Original Research

Abstract

Purpose: To develop computer-aided diagnosis (CADx) models using both mammographic and sonographic descriptors and to estimate the generalization performance of these models on future cases.

Materials and Methods: Institutional Review Board approval was obtained for this HIPPA-compliant study. Mammographic and sonographic exams were performed on 737 patients, yielding 803 breast mass lesions (296 malignant, 507 benign). Radiologist-interpreted features from the mammograms and sonograms were used as input features by a linear discriminant analysis (LDA) and an artificial neural network (ANN) to differentiate benign from malignant lesions. An LDA using all the features was compared to an LDA using only stepwise-selected features. Classification performances were quantified using receiver operating characteristic (ROC) analysis and were evaluated in a train, validate, and retest scheme. On the retest set, both LDAs were compared to the radiologists' overall assessment score of malignancy.

Results: Both the LDA and ANN achieved high classification performance with cross-validation ($AUC = 0.92 \pm 0.01$ and $_{0.90}AUC = 0.54 \pm 0.08$ for the LDA, $AUC = 0.92 \pm 0.01$ and $_{0.90}AUC = 0.55 \pm 0.08$ for the ANN). Both models also generalized very well to the re-test set, with no statistically significant performance differences between the validate and retest sets ($p > 0.1$). On the retest set, there were also no statistically significant performance differences between the LDA using all features and using only the stepwise selected features ($p > 0.3$) and between either LDA and the radiologists' assessment score ($p > 0.2$).

Conclusion: The results showed that combining mammographic and sonographic descriptors in a CADx model can result in high classification and generalization performance. On the retest set, the LDA matched the radiologists' classification performance.

Introduction

Although mammography is the only modality proven to reduce the mortality due to breast cancer, it has a low specificity for benign lesions. Because of mammography's low specificity, many women undergo unnecessary breast biopsies. As many as 65-85% of breast biopsies are performed on benign lesions (1-3). Not only does unnecessary biopsy increase the cost of mammographic screening (4), but it also subjects patients to avoidable emotional and physical burdens.

To improve the accuracy of mammography, researchers have used computer aids to help radiologists detect (5-7) and diagnose (8-11) suspicious breast lesions. Some studies have shown that such computer-aided diagnosis (CADx) systems have increased the overall diagnostic sensitivity and specificity. Lesions determined to be very likely benign may be recommended for short-term follow-up rather than biopsy (12, 13).

CADx models often use breast morphology descriptors of the Breast Imaging Reporting and Data System (BI-RADS) lexicon. BI-RADS was developed by the American College of Radiology (ACR) to standardize the interpretation of mammograms (14-17). Originally BI-RADS was applied to only mammography, but the crucial adjunct role of sonography has recently led the ACR to develop a BI-RADS lexicon for breast sonography as well. Sonographic BI-RADS is a useful tool to help standardize the characterization of sonographic lesions (17, 18) and facilitate clinician communication.

Currently, the primary clinical role for sonography is to aid in distinguishing simple cysts from solid masses, as well as to direct aspirations, wire localizations, and ultrasound guided biopsies. More recently, several authors have investigated the role of sonography in helping to differentiate malignant from benign breast lesions (19-23). There have also been many computer-aided

diagnosis studies in breast sonography, which are based upon image features automatically extracted by computer vision algorithms (24-32). To the best of our knowledge, there has not yet been a study using the standardized BI-RADS sonographic findings as the basis of a predictive model, nor to combine the use of BI-RADS mammographic and sonographic findings for that purpose.

A previous study (33) assessed the positive predictive value (PPV) and negative predictive value (NPV) of the individual sonographic BI-RADS features. This study extends previous work by using a larger database of mass lesions and by developing and evaluating decision models based upon the BI-RADS features, both mammographic and sonographic.

Materials and Methods

Patient Population

The cases for analysis in this study were an extension of the data set described in detail in a previous study (33). The cases were collected between 2000 and 2005 at our institution. The data set included 803 lesions, of which 296 were malignant and 507 were benign, and 389 were palpable and 414 nonpalpable. The patient ages ranged from 17 to 87 years, with a median age of 50 years. The same inclusion and exclusion guidelines as described previously (33) applied to this data set. Institutional review board approval was obtained for this retrospective study including a waiver of informed consent. Cases for analysis in this study were selected from those recommended for biopsy and were included in the study if the lesions corresponded to solid masses on sonography and if both mammographic and sonographic films taken before the biopsy were available for review.

Features Used

All patients underwent both mammography and sonography. The mammographic exam consisted of both craniocaudal and mediolateral-oblique views, with additional true lateral and spot compression magnification in almost all cases. Sonographic images were acquired in both radial and antiradial projections with and without caliper measurements. Additional gray-scale images were obtained in almost all cases to better show the lesion. Doppler, color Doppler, and power Doppler images were not part of the routine imaging protocol but were provided for review when available. One of four dedicated breast radiologists with 6-11 years of experience used BI-RADS lexicon descriptors to describe the lesions, as described previously (33). Information about the patient's age, physical examination findings, family history of breast cancer, and personal history of breast malignancy was available to each radiologist to most accurately reproduce a realistic clinical situation. The radiologist was blinded to the histologic diagnosis during the evaluation. Of the total 37 features, 13 were mammographic BI-RADS, 13 were sonographic BI-RADS features, 4 were ultrasound features suggested by Stavros *et al.* (19), 4 were other ultrasound features, and 3 were patient history features. The 13 mammographic BI-RADS features were mass size, parenchyma density, mass margin, mass shape, mass density, calcification number of particles, calcification distribution, calcification description, architectural distortion, associated findings, special cases, comparison with prior, and mass size. The 13 sonographic BI-RADS features were mass shape, mass orientation, mass margin, posterior acoustic features, radial diameter, antiradial diameter, anterior-posterior diameter, calcifications within mass, echo texture, lesion boundary, echo pattern, special cases, and vascularity. The five features suggested by Stavros et al were mass shape, mass margin, acoustic transmission, thin echo pesudocapsule, and mass echogenicity. The four other sonographic mass descriptors were edge shadow, cystic component, and two mammographic BI-RADS descriptors applied to ultrasound: mass shape and mass margin. The three patient history features were patient age, family history, and indication for ultrasound.

In addition to the BI-RADS and Stavros descriptors, the radiologists also recorded their assessment about the malignancy of the lesion as an integer ranging from 0 for unquestionably benign to 100 for unquestionably malignant. The gut assessment rating was not used as an input to the CADx models, but rather as a comparison to the models' output for classification performance.

Predictive Modeling and Sampling

For models in this study, we used both linear discriminant analysis (LDA) and artificial neural networks (ANNs). The LDA was a Fisher's linear discriminant. The ANNs were three-layer (one hidden layer), feed-forward, and error back-propagation artificial neural networks. These are the most popular methods used in many previous studies by our group as well as the rest of the field.

In order to assess the usefulness and risk of using computer-aided diagnosis (CADx) models in the clinic, it is crucial to have a good estimate of their performance on future cases (or generalization). For limited data and more complicated models, the traditional method of cross-validation could still pose a danger of optimistically biasing the testing performance; it is common to optimize certain global parameters (such as feature selection for the LDA or the number of hidden nodes of the ANN) to maximize cross-validation performance. With cross-validation the scientist is able to use knowledge of all the data to make modeling decisions, whereas with generalization such information is not available for yet unseen future cases. Therefore optimizing the models for cross-validation performance could lead to reduced generalization performance.

In order to avoid these overfitting pitfalls and to better estimate generalization ability of each model, we used a train, validate, and retest scheme. In this scheme the data set is divided into sets: a train/validate set and a retest set. The retest set is held aside until after the models are finalized, as not to influence any of the modeling process. All modeling decisions are made only on the train/validate set. The model parameters are optimized to maximize cross validation on the

train/validate set. Once the model's parameter values are set, the model is then trained on the entire train/validate set. The trained model is then applied to the retest set.

In particular, for our dataset of 803 lesions, we chose the first 500 cases in chronological order for the train/validate set and the remaining 303 cases for the retest set. We chose the ANN's architecture and parameter settings to optimize its cross-validation performance on the train/validate set. Once the modeling decisions had been made, we trained the LDA and ANN on all the cases in the train/validate set to determine a single, final set of weights, which were then applied to the retest set.

Classifier Performance Evaluation

To use the LDA or ANN model as a diagnostic aide, one could select a threshold value, so that cases with output values below the threshold would be considered very likely benign and therefore candidates for follow-up rather than biopsy. Those cases with model outputs greater than the threshold would be considered suspicious for malignancy and recommended for biopsy. Varying the threshold value results in a tradeoff between sensitivity and specificity. The entire range of sensitivity and specificity values for a classifier is illustrated by the receiver operating characteristic (ROC) curve (34, 35). In order to quantify a classifier's performance, we used the following five summary measures of the ROC curve: area under the ROC curve (AUC), the partial area, ($_{0.90}$AUC), as well as the specificity, positive predictive value (PPV), and negative predictive value (NPV) for a given sensitivity level. The AUC represents the average specificity over all sensitivities and ranges from 0.5 (chance performance) to 1.0 (perfect performance). Since high sensitivity is essential for a classification task, a more relevant performance measure is the $_{0.90}$AUC, which represents the average specificity performance of the classifier at sensitivities from 90% to 100%. Whereas the two previous measures provide an overall summary of performance, the remaining three are clinically relevant measures corresponding to a single threshold value,

which for breast cancer applications is usually chosen to deliver nearly perfect sensitivity such as 98% (36, 37).

Results

Generalization between Validating and Retesting

Table 1 shows the LDA performances with both 100-fold cross-validation on the train/validate set and retest performance on the retest set. The LDA achieved high classification performance, with AUC = 0.92 ± 0.01 and $_{0.90}$AUC = 0.54 ± 0.08 on the validate set and AUC = 0.92 ± 0.02 and $_{0.90}$AUC = 0.52 ± 0.08 on the retest set. The LDA generalized well; there were no statistically significant differences between the performance metrics of the validate set and those of the retest set ($p > 0.10$). In addition to the entire ROC curves of the LDA performance, individual thresholds also generalized very well. Table 2 shows that the same threshold value determined very similar true-positive fraction (sensitivity) and false-positive fraction (1-specificity) operating points in the high-sensitivity region on both ROC curves.

The ANN also performed very well, achieving AUC = 0.92 ± 0.01 and $_{0.90}$AUC = 0.55 ± 0.08 on the validate set and AUC = 0.91 ± 0.02 and $_{0.90}$AUC = 0.57 ± 0.06 on the retest set. The ANN performed comparably on the validate and retest set, with no significant differences in either metric ($p > 0.10$).

Comparison of LDA and ANN Performances

The two types of models, LDA and ANN, had very similar performances on both the validation and retest sets; the differences were not statistically significant ($p > 0.10$). In the interest of brevity, the ANN performance tables are not shown because they show very similar trends as the LDA performance tables.

Figure 1 depicts the four models' good generalization performance graphically. The ROC curves for the LDA and ANN in both testing paradigms appear in Figure 1. Figure 1 (a) shows the entire ROC curves, while 1 (b) shows only the high-sensitivity region (TPF ≥ 0.90) of those curves. The discrepancies among the curves were very minor, and the curves overlap each other. The similarity of the ROC curves showed that all four had essentially indistinguishable classification performance. Figure 1 shows good evidence of generalization for the LDA and ANN because there was no performance drop from the validation curves to the retest curves.

Feature Selection and Generalization of Simplified Model

For the LDA, we also performed a stepwise feature selection, which chose the following 14 features: patient age, calcification distribution, calcification description, associated findings, comparison with prior, anterior-posterior diameter, indication for ultrasound, Stavros mass shape, BI-RADS mass margin, edge shadow, cystic component, ultrasound lesion boundary, surrounding tissue effects, and ultrasound special findings. Feature selection was done using the validate set only. On the retest set, an LDA using only these stepwise-selected features performed comparably with no significant difference compared to the LDA using all the features (AUC = 0.92±0.02 vs. 0.91±0.02, $p > 0.3$). The full performance table for the LDA with the stepwise-selected features is not shown due to its close similarity to the table of the fully featured LDA.

Comparing the LDA to the Radiologists' Assessment of Malignancy

Table 3 compares the retest performance of the LDA against the radiologists' assessment rating on the retest set. Like the LDA, the radiologists' gut assessment also achieved high classification performance on the retest set, with AUC = 0.92 ± 0.02 and $_{0.90}$AUC = 0.52 ± 0.06 on the retest set. There were no statistically significant differences in any of the performance metrics of the LDA and radiologists' overall gut assessment ($p > 0.2$). For example, on this retest data set the LDA and radiologists performed with very similar NPV values (97±1% versus 98±1%, $p = 0.25$).

Figure 2 shows the ROC curves for the LDA with all features, the LDA with the stepwise-selected features, and the radiologists' assessment of malignancy. There were no statistically significant differences in any of the performance metrics among the three ROC curves ($p > 0.2$). Although the radiologist curve crossed over the LDA curves several times, even at the points of greater divergence, the differences were not statistically significant ($p > 0.2$).

Figure 3 depicts the histograms of the LDA output (Fig. 3 a) and radiologists' gut assessment (Fig. 3 b) values for the retest set. The histograms show the distinction in the output distributions between the benign and malignant lesions. The values for the benign lesions tended to fall on the left of the histogram plot with values around zero. Those for the malignant lesions were concentrated on the right of the plots, around one for the LDA and 100 for the radiologists' assessment values. There were few values in the center regions, compared to those on the extremes.

Example patient cases are presented in Figures 4 through 6 to illustrate situations where radiologists and computer models agree as well as disagree. Shown in Figure 4, Patient 1 presented with a well-defined, oval, well-circumscribed mass, which indicated a benign lesion. The histopathology result indicated fibroadenoma. Both the LDA and radiologist considered this case very benign, giving scores of 0.02/1.00 and 0/100, respectively. Shown in Figure 5, Patient 2 presented with a mass with irregular shape, indistinct margin, and shadowing with echogenic tails. Histopathologic diagnosis indicated that this lesion was invasive ductal carcinoma. Both the LDA and radiologist considered this case very malignant, with scores of 0.99/1.00 and 95/100, respectively. Shown in Figure 6, Patient 3 presented with a mass with an ill-defined margin in the mammogram. In the ultrasound image the lesion appeared circumscribed and oval with thick margins. Histopathologic diagnosis indicated that this lesion was necrotic breast tissue. Although some necroses could indicate malignancy, follow-up exams have shown that cancer has not appeared in this patient since biopsy two years ago. The LDA considered this case relatively

benign with a score of 0.33/1.00, whereas the radiologist considered it more indicative of malignancy with a score of 85/100.

Discussion

Previous studies have shown that BI-RADS descriptors for both mammography (4, 38-41) and sonography (19, 20, 42) are useful in predicting the likelihood of breast cancer. A previous study (33) showed that mammographic and sonographic BI-RADS features as well as Stavros ultrasound features (19) could differentiate malignant from benign breast masses with high statistical significance. Both mammographic (43, 44), and sonographic (25, 27, 45) features have been useful in breast cancer computer-aided diagnosis (CADx) systems as well. Whereas previous studies have used other features extracted from the sonogram image, to the best of our knowledge this current study is the first CADx study not only to use sonographic BI-RADS features but also the first to combine BI-RADS of ultrasound and of mammography.

In order to justify the clinical use of a CADx system on new cases, it is important to estimate its generalization performance. We have estimated the generalization performance of both an LDA and an ANN on our data set by using a train-validate-retest testing scheme on our data set. This is a more rigorous standard than most studies that rely upon train-validate only, also known as cross-validation.

The LDA and ANN had virtually indistinguishable classification performance, which indicated that the BI-RADS data were highly linear. In general, such results would support the use of the LDA model, which is simpler than the nonlinear ANN and therefore less likely to be susceptible to overtraining problems. In this study, however, it was demonstrated that there were no problems with overtraining, as both models performed very similarly during the retesting phase.

In addition to the whole ROC curve, it is important to consider more clinically relevant threshold values in determining the generalization and stability of a CADx system. Since CADx systems typically give as output a range of values, applying a certain threshold to the output determines the operating point (sensitivity and specificity settings) at which the clinical decision is made. Knowing the CADx operating point helps the clinician to incorporate it into an overall diagnostic decision. Table 2 showed that the LDA thresholds from the validation ROC curve generalized very well to the retest ROC curve in the clinically important high-sensitivity region. The threshold stability suggests that these threshold values could be used clinically with the LDA on future cases.

Because the task of collecting many features is quite cumbersome for the radiologists involved, we investigated CADx performance using only a subset of the features by performing stepwise feature selection. Of the 14 selected features, three had also been found to have high malignancy predictive value from a previous study (33): Stavros mass shape, mammographic mass margin, and sonographic lesion boundary. To assure that the selected features were adequate to allow the CADx system to generalize well on new cases, a train-test-retest scheme was required. Only the train/validate set was used to select the features, which were then tested in a CADx model on the retest set. As shown in Figure 2, an LDA with only the 14 stepwise-selected features performed just as well as an LDA with all 37 features. The small number of features required for good performance suggests that this CADx model may be able to offer the benefit of a second reader to a clinician without greatly slowing the clinician's workflow.

Figure 2 also shows that the LDA distinguished benign from malignant lesions no differently than did the radiologists' assessment scores for our data set. Note that for this data set, the actual positive predictive value of the clinical decision to refer to biopsy was 37%, which is typical of this institution. Also, since this study included only biopsy-verified cases, over this special population the sensitivity for cancer detection is by definition 100% and the specificity is 0%. The results of

this study suggest that the radiologists may be able to achieve considerable improvements in performance, such as 52% specificity, 60% PPV, and 98% NPV by adjusting their mental threshold to reduce their sensitivity slightly to 98% sensitivity, i.e., resulting in the delayed diagnosis of 2% of actual cancers which may be identified by interval change at a short-term follow-up diagnostic study. Likewise, if the radiologists were hypothetically to adopt all the recommendations of the computer model, they could have perhaps attained 37% specificity, 53% PPV, and 97% NPV at that same 98% sensitivity level.

The radiologists in this study were experienced dedicated breast imagers. It is hoped that less specialized radiologists using such a system could improve their diagnostic performance closer to that of breast specialists. In practice, it remains to be determined how radiologists would use the results from such computer models, in particular whether they would modify their biopsy recommendation in order to refer to short-term follow-up those cases deemed to be very likely benign. It also remains unknown whether the 2% of cancers mistakenly referred to follow up would prove to remain early stage such as with the current clinical practice of following probably benign cases.

As described in a previous study using this data set (33), this study's weaknesses with the BI-RADS data collection included the possibility of multiple lesions per patient, the limitation to solid masses rather than cysts, and the inclusion of only biopsy-proven lesions in the study. Additionally, radiologists allowed the mammogram to influence their recording of the sonographic features, because they analyzed the mammogram immediately before the sonogram. The study was organized in this manner to better reflect actual clinical practice in which the mammogram is obtained immediately prior to the sonogram and decision are made using all available data. They also could have shifted their diagnostic sensitivity and specificity levels from their usual clinical levels because they were aware that the cases had been resolved and therefore their assessment ratings did not directly affect patient care.

In conclusion, the models' good classification and generalization performance on our data set suggest that the models could be used as a computer-aided diagnosis (CADx) system for future mass lesions. Since the LDA threshold values generalized well, the desired operating point on the ROC curve could be set for future cases, increasing the usefulness of the CADx system. Because the stepwise-selected features were adequate for good classification and generalization, they could be used in a CADx system that would require only minimal feature collection and burden on the clinician's workflow. In this study we were not trying to improve diagnostic accuracy of dedicated breast imagers, but rather we hope to offer a tool to radiologists specializing in other specialties that will allow a substantial decrease in the number of unnecessary benign breast biopsies will minimizing the number of delayed breast cancer diagnoses.

Acknowledgments

This work was supported by US Army Breast Cancer Research Program W81XWH-05-1-0292 and DAMD17-02-1-0373, and NIH/NCI R01 CA95061 and R21 CA93461. We thank Brian Harrawood for the ROC bootstrap code, John Zhang, M.D., Carey Floyd Jr., Ph.D., Anna Bilska-Wolak, Ph.D., and Georgia Tourassi, Ph.D., for insightful discussions, and Andrea Hong, M.D., Jennifer Nicholas, M.D., Priscilla Chyn, and Susan Lim for data collection.

Tables

Table 1: Generalization of the LDA ROC Curve

Performance measure	Cross validation on train/validate set	Retest on retest set	p-value for difference in means
AUC	0.92 ± 0.01	0.92 ± 0.02	0.81
$_{0.90}$AUC	0.54 ± 0.08	0.52 ± 0.08	0.87
Spec at 98% sens	0.34 ± 0.13	0.37 ± 0.10	0.89
PPV at 98% sens	0.44 ± 0.06	0.53 ± 0.05	0.21
NPV at 98% sens	0.97 ± 0.02	0.97 ± 0.01	0.92

Caption: Table 1 shows the LDA's generalization by comparing the LDA's classification performance for 100-fold cross validation on the train/validate set (the original 500 cases) to the performance on the retest set (the latest 303 cases). The first column contains the various ROC performance metrics, whereas the LDA's score on these metrics are appears in column 2 for the train/validate set and in column 3 for the retest set. The values are shown at the mean plus or minus one standard deviation, as determined by bootstrap analysis of the ROC curves. The last column shows the two-tailed p-values for the difference in the two sets' performance metric values, as determined by two-sided *t* tests. The LDA achieved high classification performance. Since there were no statistically significant differences between the performance metrics, the LDA performed equivalently on cross validating on the train/validate set and on retesting on the retest set: The LDA generalized well.

Table 2: Generalization of the LDA Threshold

LDA Threshold	TPF on Validate ROC	TPF on Retest ROC	FPF on Validate ROC	FPF on Retest ROC
0.0782	0.953	0.945	0.429	0.466
0.0373	0.976	0.976	0.613	0.591
0.0201	0.982	0.984	0.728	0.739
0.0098	1	1	0.879	0.886

Caption: The LDA thresholds from the validation ROC curve generalized very well to the retest ROC curve. The same threshold value determined very similar true-positive fraction (sensitivity) and false-positive fraction (1-specificity) operating points on both ROC curves. Such performance stability is clinically important for computer-aided diagnosis (CADx) systems; knowing the CADx operating point helps the clinician to incorporate it into an overall diagnostic decision.

Table 3: LDA vs. Radiologists' Overall Gut Assessment on the Retest Set

Performance measure	LDA	Radiologists' overall gut assessment	p-value for difference in means
AUC	0.92 ± 0.02	0.92 ± 0.02	0.98
$_{0.90}$AUC	0.52 ± 0.08	0.52 ± 0.06	0.98
Spec at 98% sens	0.37 ± 0.10	0.52 ± 0.08	0.25
PPV at 98% sens	0.53 ± 0.05	0.60 ± 0.05	0.25
NPV at 98% sens	0.97 ± 0.01	0.98 ± 0.01	0.25

Caption: The table compares the LDA to the radiologists' overall gut assessment on the retest set. Column 1 lists the ROC performance metrics, column 2 the LDA's performance, column 3 the radiologists' performance, and column 4 the two-tailed p-value for the difference in means. The p-values and errors on the classification performance metric values were determined by ROC bootstrap analysis. Both the LDA and the radiologists achieved excellent classification performance and performed equivalently, with no statistically significant performance differences between them.

Figures

Figure 1 a: Full ROC Curves: Validation vs. Retest

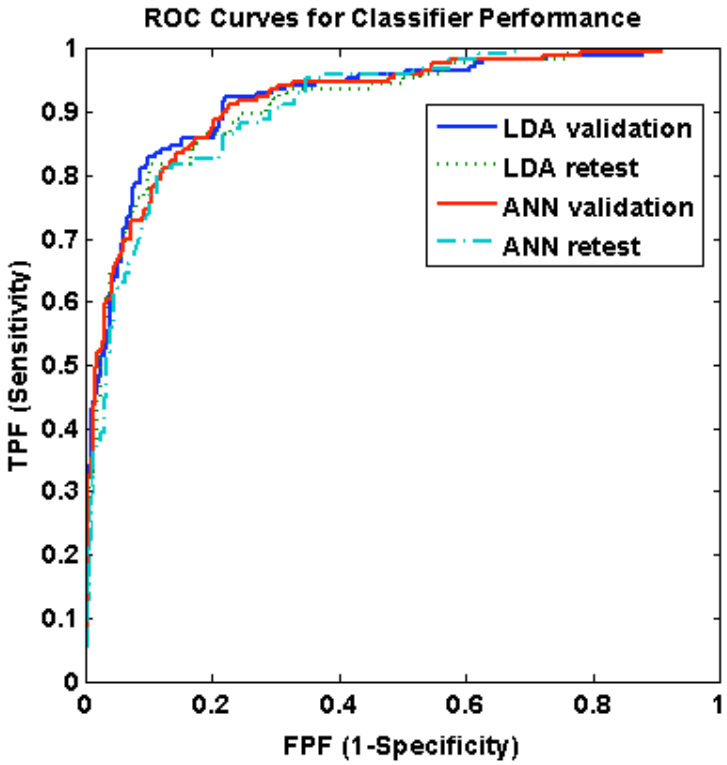


Figure 1 b: Partial ROC Curves: Cross Validation vs. Retest

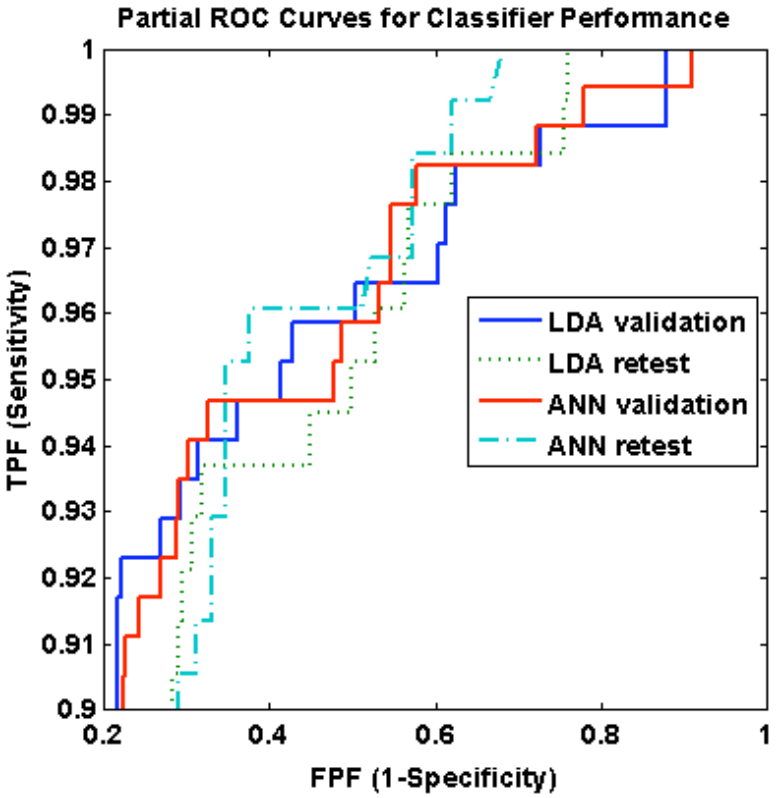


Figure 1 Caption: Both the LDA and the ANN generalized well on the retest data set, as shown by their overlapping ROC curves. The validation ROC curves (solid curves) lie very close to the retest ROC curves (dashed curves). The LDA and ANN had virtually indistinguishable classification performances.

Figure 2 a: Full ROC Curves: LDA vs. Radiologist, Retest Set

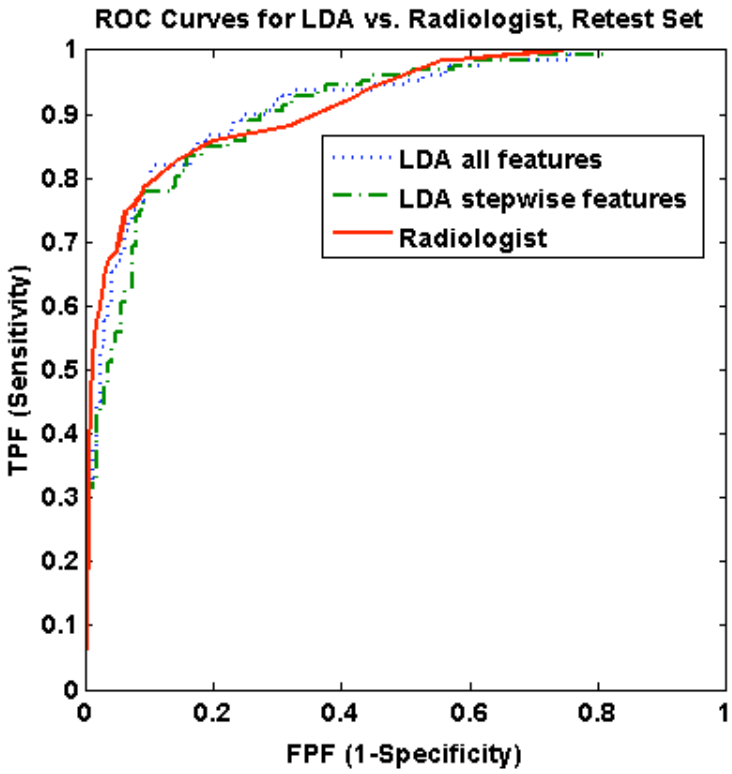


Figure 2 b: Partial ROC Curves: LDA vs. Radiologist, Retest Set

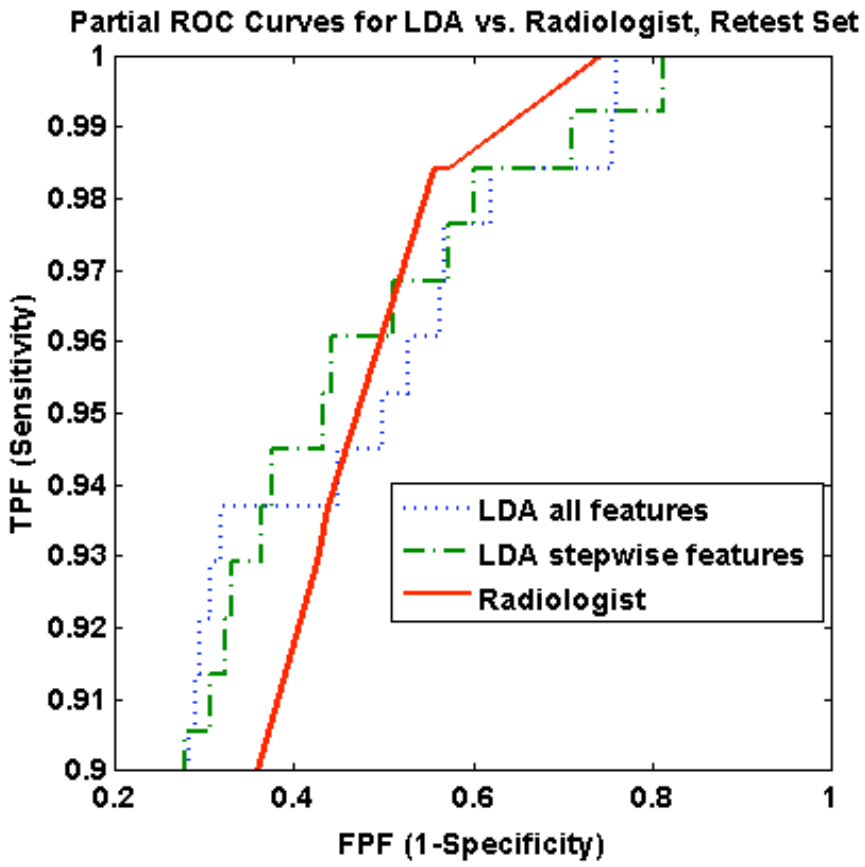


Figure 2 Caption: Shown here are the ROC curves for the LDA with all features, for the LDA with the stepwise-selected features, and for the radiologists' assessment of malignancy. In retesting, the LDA, both using all features and using the stepwise-selected features, performed very similarly to the radiologists' overall gut assessment scoring. There were no statistically significant differences in any of the performance metrics among the three ROC curves ($p > 0.2$). Although the radiologist curve crossed over the LDA curves several times, even at the points of greater divergence, the differences were not statistically significant ($p > 0.2$).

Figure 3 a: Histograms of the LDA Output Values

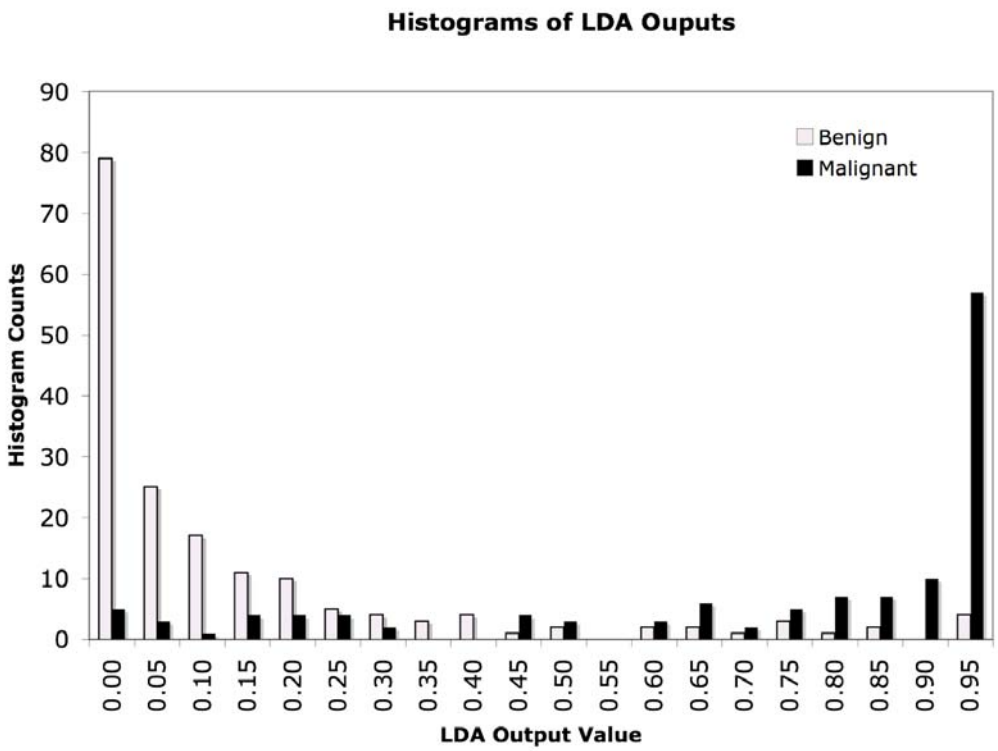


Figure 3 b: Histograms of the Radiologists' Overall Gut Assessment

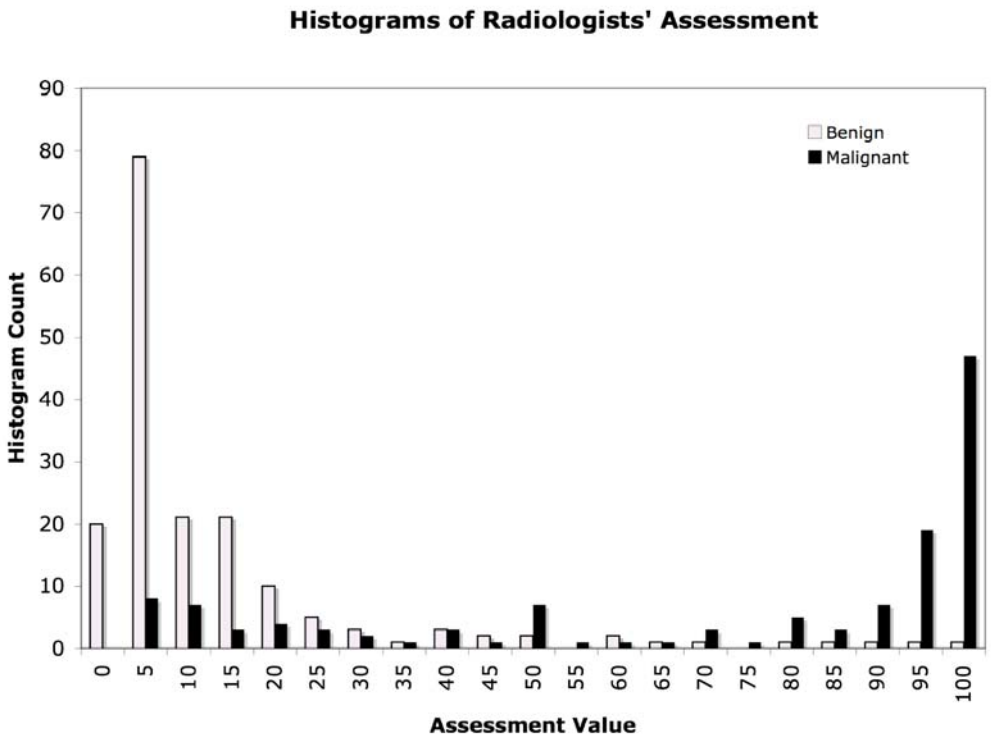


Figure 3 Caption: Plotted above are histograms of the LDA output values (a) and of the radiologists' overall gut assessment values (b). The histogram counts for the truly benign lesions are shown in gray, and those for the truly malignant lesions are shown in black. For classification, a threshold would be applied to the LDA output, so that output values below the threshold would be designated benign and those above it would be designated malignant.

Figure 4 a: Mammogram of Patient 1

Figure 4a Caption: Mediolateral oblique mammographic view in 52 year-old woman demonstrates an oval, well-circumscribed, equal density mass in the superior left breast.

Figure 4 b: Sonogram of Patient 1

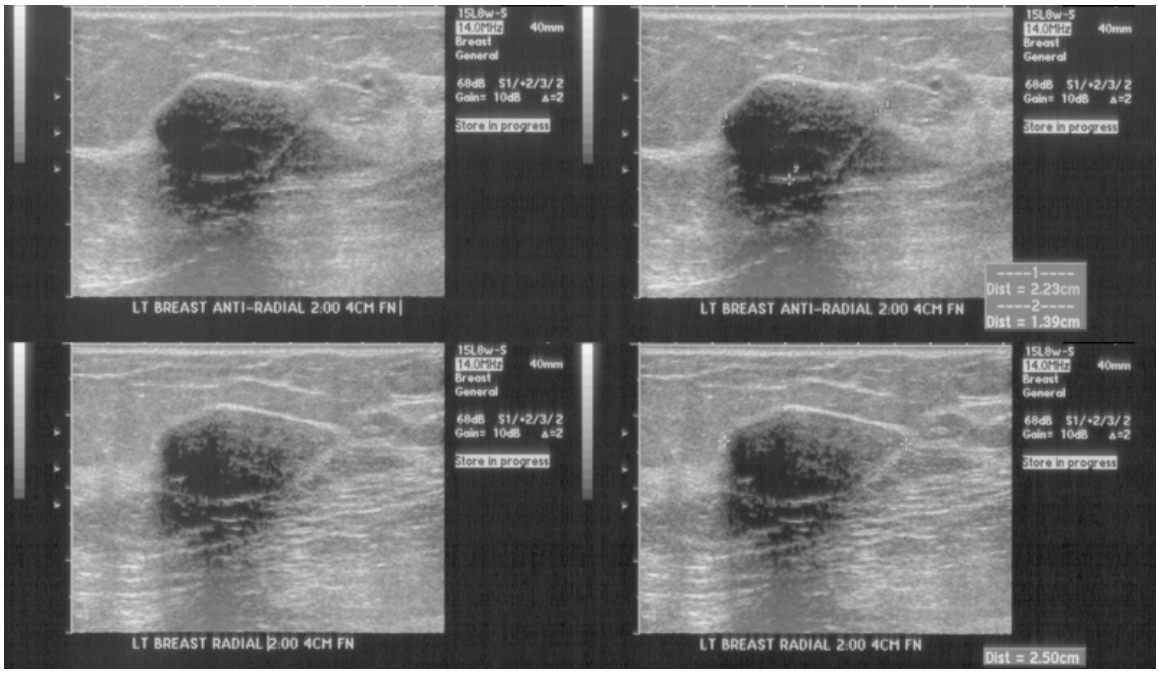


Figure 4b Caption: Ultrasound views of the mass demonstrate an oval, hypoechoic solid mass with circumscribed margins, parallel orientation and posterior acoustic shadowing The histopathology result indicated a benign fibroadenoma. Both the LDA and radiologist correctly considered this case very benign, giving scores of 0.02/1.00 and 0/100, respectively.

Figure 5 a: Mammogram of Patient 2

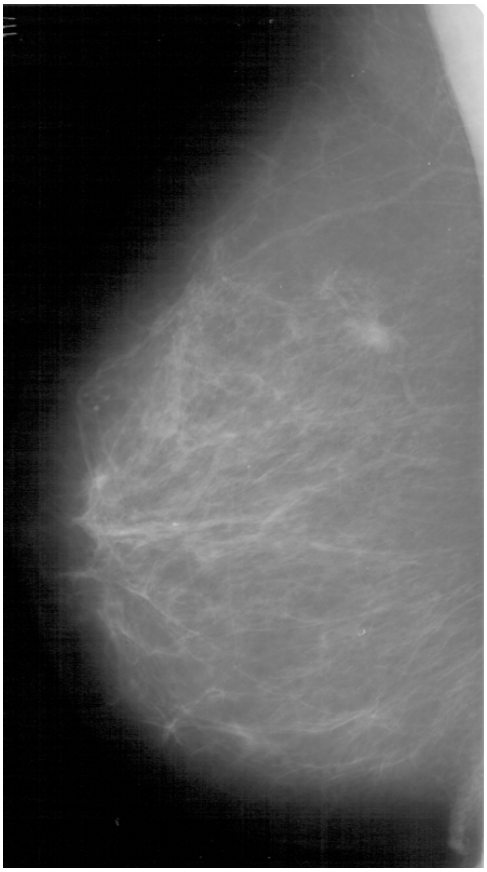


Figure 5a Caption: Mediolateral oblique mammographic view in 57 year-old woman demonstrates an ill-defined, irregularly-shaped, equal density mass in the superior right breast.

Figure 5 b: Sonogram of Patient 2

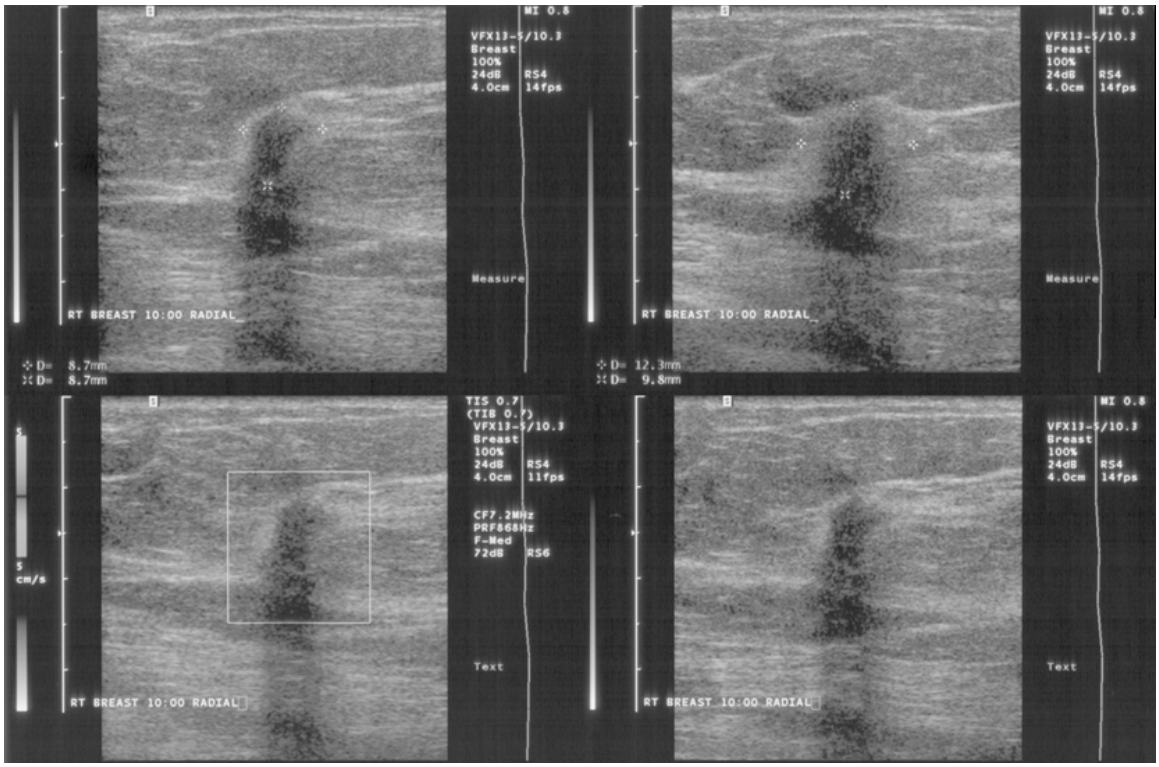


Figure 5b Caption: US views of the mass demonstrate an ill-defined, irregularly-shaped mass with posterior acoustic shadowing and not-parallel orientation. Histopathologic diagnosis indicated that this malignant lesion was invasive ductal carcinoma. Both the LDA and radiologist correctly considered this case very malignant, with scores of 0.99/1.00 and 95/100, respectively.

Figure 6 a: Mammogram of Patient 3

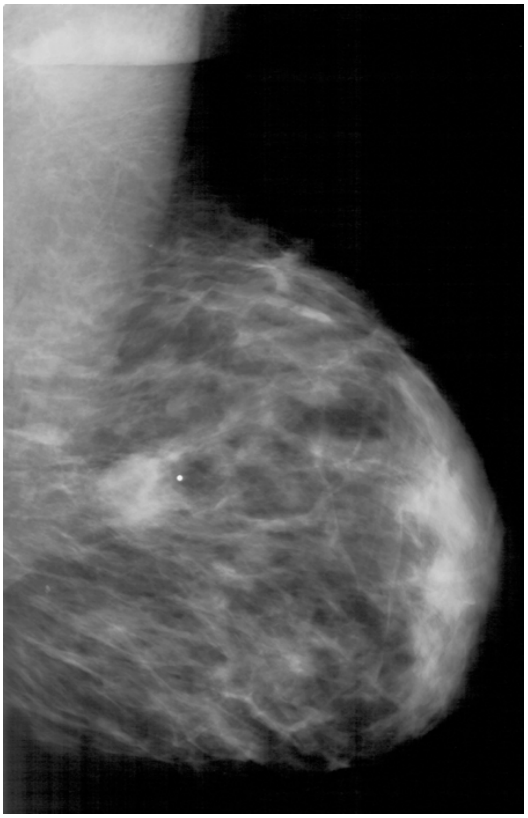


Figure 6a Caption: Mediolateral oblique mammographic view in 26 year-old woman demonstrates an ill-defined, oval-shape, equal density mass in the posterior left breast.

Figure 6 b: Sonogram of Patient 3

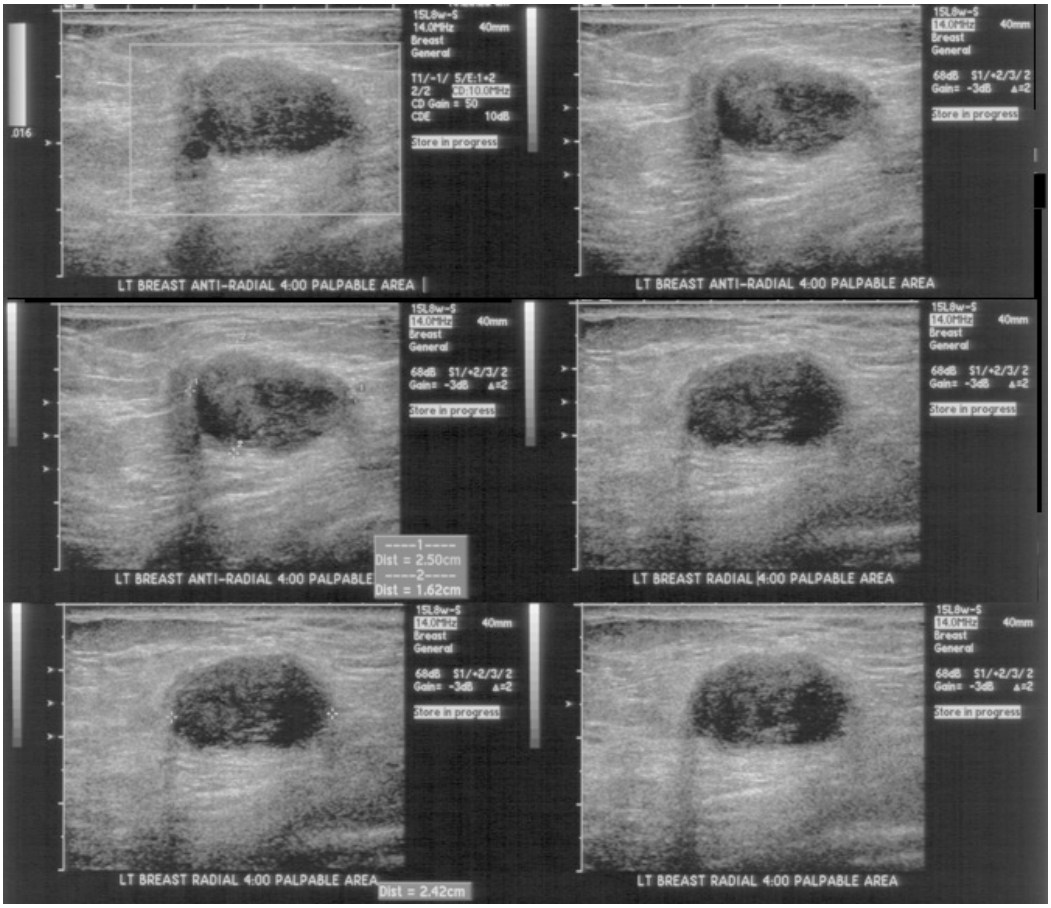


Figure 6 Caption: Sonographic views of the mass demonstrate an oval, circumscribed mass with parallel orientation and no posterior acoustic features. Histopathologic diagnosis indicated that this lesion was necrotic breast tissue. Follow-up exams confirm no interval change two years post biopsy. The LDA considered this case relatively benign with a score of 0.33/1.00, whereas the radiologist considered it more indicative of malignancy with a score of 85/100.

References

1. Kopans DB. The Positive Predictive Value of Mammography. American Journal of Roentgenology 1992; 158:521-526.
2. Ciatto S, Cataliotti L, Distante V. Nonpalpable lesions detected with mammography: review of 512 consecutive cases. Radiology 1987; 165:99-102.
3. Meyer JE, Eberlein TJ, Stomper PC, Sonnenfeld MR. Biopsy of occult breast lesions. Analysis of 1261 abnormalities. JAMA 1990; 263:2341-2343.
4. Cyrlak D. Induced costs of low-cost screening mammography. Radiology 1988; 168:661-663.
5. Zheng B, Chang YH, Wang XH, Good WF, Gur D. Feature selection for computerized mass detection in digitized mammograms by using a genetic algorithm. Acad. Radiol. 1999; 6:327-332.
6. Qian W, Clarke LP, Song D, Clark RA. Digital mammography: hybrid four-channel wavelet transform for microcalcification segmentation. Acad. Radiol. 1998; 5:354-364.
7. Qian W, Li L, Clarke L, Clark RA, Thomas J. Digital mammography: comparison of adaptive and nonadaptive CAD methods for mass detection. Acad. Radiol. 1999; 6:471-480.
8. Chan HP, Sahiner B, Helvie MA, et al. Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999; 212:817-827.
9. Chan HP, Sahiner B, Lam KL, et al. Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces. Med. Phys. 1998; 25:2007-2019.
10. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad. Radiol. 1999; 6:22-33.
11. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad. Radiol. 1998; 5:155-168.
12. Baker JA, Kornguth PJ, Lo JY, Williford ME, Floyd CE, Jr. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. Radiology 1995; 196:817-822.
13. Lo JY, Baker JA, Kornguth PJ, Floyd CE, Jr. Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks. Acad. Radiol. 1999; 6:10-15.
14. Kopans DB. Standardized mammography reporting. Radiologic Clinics of North America 1992; 30:257-264.
15. D'Orsi CJ, Kopans DB. Mammographic feature analysis. Seminars in Roentgenology 1993; 28:204-230.

16. BI-RADS. American College of Radiology Breast Imaging - Reporting and Data System (BI-RADS) 3rd ed. In. Reston, VA: American College of Radiology, 1998.
17. Radiology ACo. American College of Radiology. BI-RADS: ultrasound, 1st ed. In: Breast Imaging - Reporting and Data System (BI-RADS) atlas, 4th ed. . Reston, VA, 2003.
18. Mendelson EB, Berg WA, Merritt CR. Toward a standardized breast ultrasound lexicon, BI-RADS: ultrasound. Seminars in Roentgenology 2001; 36:217-225.
19. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker S, Sisney G. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. Radiology 1995; 196:123-134.
20. Rahbar G, Sie AC, Hansen GC, et al. Benign versus malignant solid breast masses: US differentiation. Radiology 1999; 213:889-894.
21. Jackson VP. The role of US in breast imaging. Radiology 1990; 177:305-311.
22. Jackson VP. Management of solid breast nodules: what is the role of sonography? Radiology 1995; 196:14-15.
23. Zonderland HM, Coerkamp EG, Hermans J, van de Vijver MJ, van Voorthuisen AE. Diagnosis of breast cancer: contribution of US as an adjunct to mammography. Radiology 1999; 213:413-422.
24. Chang RF, Kuo WJ, Chen DR, Huang YL, Lee JH, Chou YH. Computer-aided diagnosis for surgical office-based breast ultrasound. Archives of Surgery 2000; 135:696-699.
25. Chen D, Chang RF, Huang YL. Breast cancer diagnosis using self-organizing map for sonography. Ultrasound Med. Biol. 2000; 26:405-411.
26. Giger ML. Computerized analysis of images in the detection and diagnosis of breast cancer. Semin Ultrasound CT MR 2004; 25:411-418.
27. Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. Acad Radiol 2004; 11:272-280.
28. Drukker K, Giger ML, Vyborny CJ, Mendelson EB. Computerized detection and classification of cancer on breast ultrasound. Acad Radiol 2004; 11:526-535.
29. Drukker K, Horsch K, Giger ML. Multimodality computerized diagnosis of breast lesions using mammography and sonography. Acad. Radiol. 2005; 12:970-979.
30. Drukker K, Giger ML, Metz CE. Robustness of computerized lesion detection and classification scheme across different breast US platforms. Radiology 2005; 237:834-840.
31. Moon WK, Chang RF, Chen CJ, Chen DR, Chen WL. Solid breast masses: classification with computer-aided analysis of continuous US images obtained with probe compression. Radiology 2005; 236:458-464.
32. Chen DR, Chang RF, Chen CJ, et al. Classification of breast ultrasound images using fractal feature. Clinical Imaging 2005; 29:235-245.
33. Hong AS, Rosen EL, Soo MS, Baker JA. BI-RADS for Sonography: Positive and Negative Predictive Values of Sonographic Features. Am. J. Roentgenol. 2005; 184:1260-1265.
34. Metz CE. Basic principles of ROC analysis. Sem Nuc Med 1978; 8:283-298.

35. Metz C. Evaluation of CAD methods. In: Doi K, MacMahon H, Giger ML, Hoffmann KR, eds. Computer-aided Diagnosis in Medical Imaging. Amsterdam: Elsevier Science, 1998; 543-554.
36. Zhou X-H, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York, NY: John Wily & Sons, 2002.
37. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996; 201:745-750.
38. Hall FM, Storella JM, Silverstone DZ, Wyshak G. Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography. Radiology 1988; 167:353-358.
39. Chang YH, Hardesty LA, Hakim CM, et al. Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment. Med. Phys. 2001; 28:455-461.
40. Obenauer S, Hermann KP, Grabbe E. Applications and literature review of the BI-RADS classification. European Radiology 2005; 15:1027-1036.
41. Lee S, Lo C, Wang C, et al. A computer-aided design mammography screening system for detection and classification of microcalcifications. International Journal of Medical Informatics 2000; 60:29-57.
42. Harper AP, Kelly-Fry E, Noe JS, Bies JR, Jackson VP. Ultrasound in the evaluation of solid breast masses. Radiology 1983; 146:731-736.
43. Chan HP, Sahiner B, Petrick N, et al. Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network. Phys. Med. Biol. 1997; 42:549-567.
44. Petrick N, Sahiner B, Chan HP, Helvie MA, Paquerault S, Hadjiiski LM. Breast cancer detection: Evaluation of a mass-detection algorithm for computer-aided diagnosis - Experience in 263 patients. Radiology 2002; 224:217-224.
45. Horsch K, Giger ML, Venta LA, Vyborny CJ. Computerized diagnosis of breast lesions on ultrasound. Med. Phys. 2002; 29:157-164.

The Effect of Data Set Size on Computer-Aided Diagnosis of Breast Cancer: Comparing Decision Fusion to a Linear Discriminant

Jonathan L. Jesneck[1,2], Loren W. Nolte[1,3], Jay A. Baker[2], Joseph Y. Lo[1,2]

[1] Department of Biomedical Engineering, Duke University, Durham, NC 27708
[2] Duke Advanced Imaging Laboratories, Department of Radiology,
2424 Erwin Road, Suite 302, Durham, NC 27705
[3] Department of Electrical and Computer Engineering, Duke University, Durham, NC 27705

ABSTRACT

Data sets with relatively few observations (cases) in medical research are common, especially if the data are expensive or difficult to collect. Such small sample sizes usually do not provide enough information for computer models to learn data patterns well enough for good prediction and generalization. As a model that may be able to maintain good classification performance in the presence of limited data, we used decision fusion. In this study, we investigated the effect of sample size on the generalization ability of both linear discriminant analysis (LDA) and decision fusion. Subsets of large data sets were selected by a bootstrap sampling method, which allowed us to estimate the mean and standard deviation of the classification performance as a function of data set size. We applied the models to two breast cancer data sets and compared the models using receiver operating characteristic (ROC) analysis. For the more challenging calcification data set, decision fusion reached its maximum classification performance of AUC = 0.80±0.04 at 50 samples and pAUC = 0.34±0.05 at 100 samples. The LDA reached a lower performance and required many more cases, with a maximum of AUC = 0.68±0.04 and pAUC = 0.12±0.05 at 450 samples. For the mass data set, the two classifiers had more similar performance, with AUC = 0.92±0.02 and pAUC = 0.48±0.02 at 50 samples for decision fusion and AUC = 0.92±0.03 and pAUC = 0.55±0.04 at 500 samples for the LDA.

Keywords: Decision Fusion, Computer-Aided Diagnosis, Sample Size, Receiver Operating Characteristic (ROC) Curve, Classification, Breast Cancer

1. INTRODUCTION

Many medical data sets are difficult and expensive to collect, often resulting in limited data set size. A small number of cases usually precludes accurate predictive modeling. Early modeling offers many advantages, such as earlier identification of data collection problems, of unsatisfactory patient sampling, of expensive but uninformative features, and perhaps earlier discovery of flaws in the scientific experiment design. Many medical experiments expose subjects to possibly avoidable risk that could be detected by better and earlier modeling.

The amount of available data affects each model differently. Model complexity tends to produce a tradeoff between modeling power and generalization; simpler models may be more robust to noise in the data but may not be able to capture the full complexity of the data's patterns, whereas more complicated models may model the patterns better but are more susceptible to overfitting. In addition to the number of samples available, the ratio of number of features to number of samples can also affect classifier performance. Many classical models tend to overtrain on data sets with few samples and many features. This overtraining effect becomes more pronounced with smaller sample size.

In this study, we investigated the effect of sample size on the generalization ability of two computer-aided diagnosis (CADx) models. The first model was linear discriminant analysis (LDA), a common CADx model for breast cancer data. The second model was a decision-fusion method that has shown promise for small, noisy data sets[1]. Our decision-fusion technique offers the significant advantage that it can reduce the dimensionality of the feature space

of the classification problem by assigning a classifier to each feature separately. Considering only one feature at a time greatly reduces the complexity of the problem by avoiding the need to estimate multidimensional probability density functions (PDFs) of the feature space. Accurately estimating multidimensional PDFs likely requires many more observations than a typical medical data set contains[2]. Considering only one-dimensional PDFs may allow the decision-fusion technique to reach asymptotic testing performance using many fewer cases than other classifiers require.

Other benefits of decision fusion are that it is robust in noisy data[3], is not overly sensitive to the likelihood ratio threshold values[4], and can handle missing data values[5]. Our decision-fusion technique can also be tuned to optimize arbitrary performance metrics that may be more clinically relevant, unlike more traditional classification algorithms that optimize mean squared error, such as the LDA.

II. METHODS

2.1 Data

This study used two breast cancer data sets: one of mass lesions and one of calcification lesions.

The mass lesion data set is an extension of the earlier subset described by Hong, *et al.* from this research group[6]. The cases were collected between 2000 and 2005 at Duke University. The data set included 803 lesions, of which 296 were malignant and 507 were benign, and 389 were palpable and 414 nonpalpable. The patient ages ranged from 17 to 87 years, with a median age of 50 years. Patients underwent both mammography and sonography, and outcome was determined through definitive histopathological diagnosis. One of three dedicated breast radiologists with 6-11 years of experience described each lesion using Breast Imaging Reporting and Data System (BI-RADS™, American College of Radiology, Reston, VA)[7] mammography, BI-RADS sonography, and Stavros sonography descriptors[6]. Of the total 38 features, 13 were mammographic, 22 were sonographic, and 3 were patient history features.

Second, we used a calcification data set that consisted of 1508 mammogram microcalcification lesions from the Digital Database for Screening Mammography (DDSM)[8], which is publicly available. The outcomes were verified by histopathological diagnosis and follow-up for certain benign cases, yielding 811 benign and 697 malignant calcification lesions. The feature groups were 13 computer-extracted calcification cluster morphological features, 91 computer-extracted texture features of the lesion background anatomy, 2 radiologist-interpreted findings, 2 radiologist-extracted features from the BI-RADS lexicon and patient age. In total, calcification data C set had 109 features and a sample-to-feature ratio of approximately 14:1. Each mammogram was digitized with a resolution of either 43.5 microns (Howtek 960 or MultiRad850 digitizer) or 50 microns (Lumisys 200 Laser digitizer). We used a 512x512 pixel ROI centered on the centroid of each lesion (using lesion outlines drawn by the DDSM radiologists) for image processing and for generating the computer-extracted features. We extracted morphological and texture (spatial gray level dependence matrix) features, which were shown to be useful in previous studies of CADx such as by Chan, *et al.*[9].

2.1 Decision Fusion

For the decision-fusion classifier, histograms of each feature were constructed as an estimate of the probability density in order to construct an empirical likelihood ratio for that feature. Then, a binary decision was made by comparing the likelihood ratio value to a given threshold, which in turn determined the sensitivity and specificity of the decision. Finally, the decision fusion theory allowed the individual binary decisions to be combined optimally to produce one final binary decision.

First, each feature was considered separately and classified by a likelihood ratio classifier. According to decision theory, the likelihood ratio is the optimal detector to determine the presence or absence of a signal in noise[10]. The null hypothesis (H_0) was that the signal is not present in the noisy features, while the alternative hypothesis (H_1) was that the signal is present.

$$H_0 : X = N$$
$$H_1 : X = S + N \tag{1}$$

The likelihood ratio is the probability of the features under the malignant case divided by the probability of the features under the benign case:

$$\lambda(X) = \frac{P(X \mid H_1)}{P(X \mid H_0)}, \tag{2}$$

where $p(X|H_1)$ is the PDF of the observation data X given that the signal is present, and $p(X|H_0)$ is the PDF of the data X given that the signal is not present. The likelihood ratio is optimal under the assumption that the PDFs accurately reflect the true densities. For classification, we can apply a threshold value, τ, to the likelihood ratio to produce a binary decision, u, about the presence of the signal.

$$u = \begin{cases} 1 & \text{if } \lambda \geq \tau \\ 0 & \text{if } \lambda < \tau \end{cases} \tag{3}$$

Since we assigned a separate likelihood ratio classifier to each of *p* features, we applied a separate threshold to each classifier's output value to produce *p* binary decisions. A genetic algorithm searched over the joint set of thresholds in order to maximize the classification performance of the fused binary decisions. The genetic algorithm search time was capped at 30 generations for this study due to computational cost.

Decision-fusion theory describes how to combine local binary decisions optimally to determine the presence or absence of a signal in noise[11-15]. The decision fuser optimally fuses all the local decisions according to the operating points on the receiver operating characteristic (ROC) curve at which the local decisions were made. Assuming statistically independent decisions, the likelihood ratio of the fused classifier is a product over the "yes, signal present" ($u_i = 1$) decisions multiplied by a similar product over the "no, signal absent" ($u_i = 0$) decisions.

$$\lambda_{fused}(u_1,\ldots,u_p) = \prod_{\substack{i \\ u=1}} \frac{Pd_i}{Pf_i} \prod_{\substack{i \\ u=0}} \frac{1-Pd_i}{1-Pf_i}, \tag{4}$$

where Pd_i is the probability of detection or sensitivity, and Pf_i is the probability of false detection, or (1-specificity), for the i^{th} local decision. The ROC curve can be computed from the unique likelihood-ratio values of the fused classifier as shown in Equation (5).

$$\begin{aligned} Pd_{fused}(j) &= \sum_{i=j}^{p} P(\lambda_{fused,i} \mid H_1), \quad j = 0,\ldots,p \\ Pf_{fused}(j) &= \sum_{i=j}^{p} P(\lambda_{fused,i} \mid H_0), \quad j = 0,\ldots,p \end{aligned} \tag{5}$$

2.2 Linear Discriminant Analysis

The baseline classifier was linear discriminant analysis (LDA), which served as a benchmark for the linear separability of the data set.

2.3 Sampling and Validation

In order to study the effect of sample size on the classifiers' performances, we randomly selected subsets of the data sets. We varied the number of selected cases from 50 to 500, which covers typical data set sizes in preliminary CADx research. Ten random draws of each data subset size were drawn to assess selection effects. On each subset, both classifiers were trained and validated using 10-fold cross-validation. For each sample size such as 100 cases, classifiers were developed using ten bootstrap samples of that number of cases, which allowed the calculation of the mean AUC and pAUC values along with their standard deviations.

2.4 Classifier Comparison

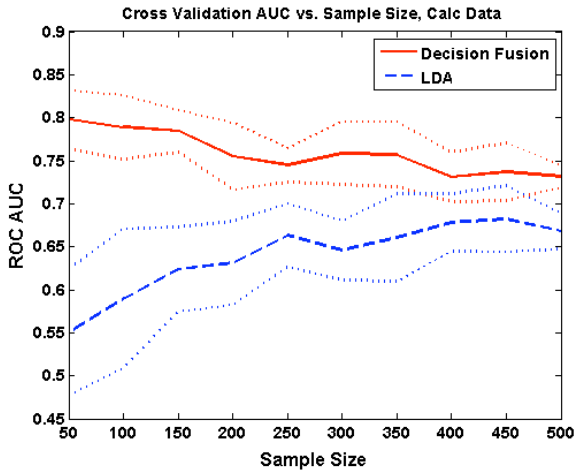
Each classifier was evaluated using ROC analysis. Two clinically interesting summary metrics of the ROC curve were used: the area under the curve (AUC) and the normalized partial area of the curve (pAUC), which is measured above sensitivity of Pd = 0.9.

III. RESULTS

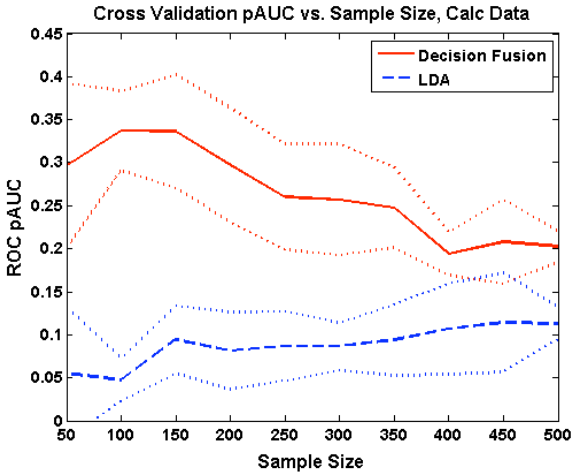
Figure 1 plots the classification performance against the number of cases. The classifiers' performances were scored both by ROC AUC (Fig. 1a and 1c) and pAUC (1b and 1d).

On the calcification data (Fig. 1a and 1b) decision fusion achieved a maximum of AUC = 0.80±0.04 at 50 samples and pAUC = 0.34±0.05 at 100 samples. The LDA had a lesser performance, with AUC = 0.68±0.04 and pAUC = 0.12±0.05 at 450 samples. The LDA had the expected testing trend of slowly increasing performance with increasing sample size, but decision fusion showed the opposite trend. Perhaps inadequately trained, decision fusion decreased with sample size both in AUC and pAUC. Note that all of these are validation results from k-fold cross-validation, which normally should minimize effects of training bias.

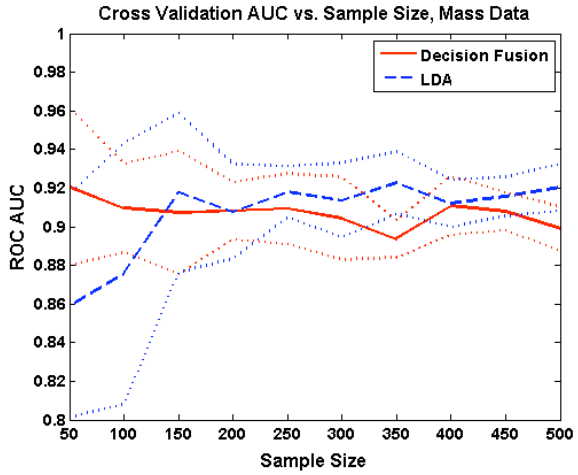
For the mass lesion data (Fig. 1b and 1d), the two classifiers' performances had more similar trends. Decision fusion reached a maximum of AUC = 0.92±0.02 and pAUC = 0.48±0.02 at 50 samples, and the LDA reached AUC = 0.92±0.03 and pAUC = 0.55±0.04 at 500 samples. No significant performance differences between the classifiers were seen in sample sizes greater than 100. For very small data sets of 50 cases, decision fusion outperformed the LDA. In both data sets, decision fusion approached its final AUC value with many fewer cases than the LDA required. All plots except Fig. 1b showed that decision fusion had a smaller slope than the LDA.



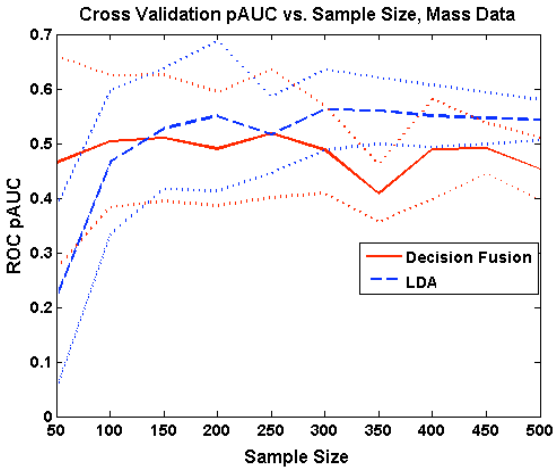
(a) AUC vs. Sample Size, Calcification Data



(b) pAUC vs. Sample Size, Calcification Data



(c) AUC vs. Sample Size, Mass Data



(d) pAUC vs. Sample Size, Mass Data

Figure 1: Classifier performance vs. Sample Size

Decision fusion significantly outperformed the LDA on the calcification data set. The performance difference was greatest for small data sets. However, on the larger data sets, the performance gap narrowed to 0.06. In part (b), decision fusion achieved pAUC = 0.34±0.05 at 100 samples and then fell to pAUC = 0.2±0.02 at 500 samples. Although the two classifiers had very similar performance on the mass data set, decision fusion still outperformed the LDA for very small sample sizes.

IV. DISCUSSION

Decision fusion had its biggest classification performance gain over the LDA on the noisier, more nonlinear data set, the calcification data set. On the mass data set, both the LDA and decision fusion performed very similarly for data sets larger than 50 samples. On very small data sets of 50 samples, which are common among initial CADx studies, decision fusion outperformed the LDA. For the mass data set at least, a particular strength of the decision-fusion algorithm is that it is able to estimate asymptotic testing performance with many fewer cases than other classifiers require. Figure 1 shows that decision fusion was able to achieve approximately the same testing performance with 50 cases as with 500 cases.

The general downward slope of the decision fusion curves for the calcification data set may be due to inadequate training. For computational convenience, we limited the genetic algorithm's search time to only 30 generations. Whereas 30 generations were adequate for small data sets smaller than 150 cases, larger data sets required more genetic algorithm generations for complete optimization. A much longer run of 3000 generations on all available 1508 cases in the calcification lesion data set improved decision fusion's performance under 100-fold cross-validation to AUC = 0.85±0.01 and pAUC = 0.28±0.03, which exceeded the performance for all data points shown in Fig. 1a and 1b. A similar more thorough optimization on all available 803 cases in the mass data set allowed decision fusion to reach AUC = 0.94±0.01 and pAUC = 0.63±0.07, which likewise also exceeded the performances in Fig. 1c and 1d.

The improvements were usually significant for the more challenging calcification data set, but not for the mass data set. Such a statement may not reflect the full diversity of these data sets, which differ in many respects, including linear separability, numbers of cases, numbers and types of features, and feature correlations. Future work will explore the contribution of such factors using controlled simulation data sets in order to understand the full potential and limitations of the decision-fusion technique.

ACKNOWLEDGEMENTS

This work was supported by US Army Breast Cancer Research Program W81XWH-05-1-0292 and DAMD17-02-1-0373, and NIH/NCI R01 CA95061 and R21 CA93461. I would also like to thank Brian Harrawood for the ROC bootstrap code, Anna Bilska-Wolak, Ph.D., and Georgia Tourassi, Ph.D., for insightful discussions, and Andrea Hong, M.D., Jennifer Nicholas, M.D., Priscilla Chyn, and Susan Lim for data collection.

REFERENCES

1. Jesneck JL, Lo JY, Baker JA. "A computer aid for diagnosis of breast mass lesions using both mammographic and sonographic descriptors" Radiology (submitted January 2006).
2. Hastie T, R. T, Friedman JH. *The Elements of Statistical Learning*, Springer, 2001.
3. Niu R, Varshney PK, Moore M, Klamer D. "Decision fusion in a wireless sensor network with a large number of sensors". International Society of Information Fusion, Fairborn, OH 45324, United States, 2004; 21.
4. Zhu M, Ding S, Brooks RR, Wu Q, Rao NSV, Iyengar SS. "Fusion of threshold rules for target detection in sensor networks". ACM Transactions on Sensors Networks (submitted for publication).
5. Bilska-Wolak AO, Floyd CE, Jr. "Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer". Phys Med Biol 2004; **49**:4219-4237.

6. Hong AS, Rosen EL, Soo MS, Baker JA. “BI-RADS for Sonography: Positive and Negative Predictive Values of Sonographic Features”. Am J Roentgenol 2005; **184**:1260-1265.
7. American College of Radiology, *Breast Imaging - Reporting and Data System (BI-RADS)* 3rd ed., Reston, VA, American College of Radiology, 1998.
8. Heath M, Bowyer KW, Kopans D. “Current status of the Digital Database for Screening Mammography”. In: *Digital Mammography*, Karssemeijer N, Thijssen M, Hendriks J, eds.: Kluwer Academic Publishers, p. 457-460, 1998.
9. Chan HP, Sahiner B, Lam KL, et al. “Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces”. Medical Physics 1998; **25**:2007-2019.
10. VanTrees HL. *Detection, Estimation, and Modulation Theory (Part I)*. John Wiley & Sons, New York, 1968.
11. Tenney RR, Sandell NR, Jr. “Detection with Distributed Sensors”. *Proc IEEE Conf Incl Symp Adapt Processes*, 1980; **1**:433.
12. Chair Z, Varshney PK. “Optimal data fusion in multiple sensor detecton systems”. *IEEE Transactions on Aerospace and Electronic Systems* 1986; AES-**22**:98.
13. Reibman AR, Nolte LW. “Optimal detection and performance of distributed sensor systems”. *IEEE Transactions on Aerospace and Electronic Systems* 1987; AES-**23**:24.
14. Dasarathy BV. “Decision fusion strategies in multisensor environments”. *IEEE Transactions on Systems, Man and Cybernetics* 1991; **21**:1140.
15. Liao Y. *Distributed decision fusion in signal detection -- a robust approach*. Ph.D. Thesis, Duke Univeristy, 2005.